Statistical Module 1; number scales and frequency distributions

- ✓ Click “slide show” to start this presentation as a show.
- ✓ **Remember: focus & think about each point; do not just passively click.**

Cranach, *Tree of Knowledge* [of Good and Evil] (1472)

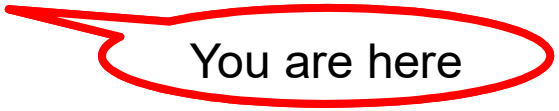
© Dr. David J. McKirnan, 2014
The University of Illinois Chicago
McKirnanUIC@gmail.com
Do not use or reproduce without permission

SHOUT
SHARE
IT

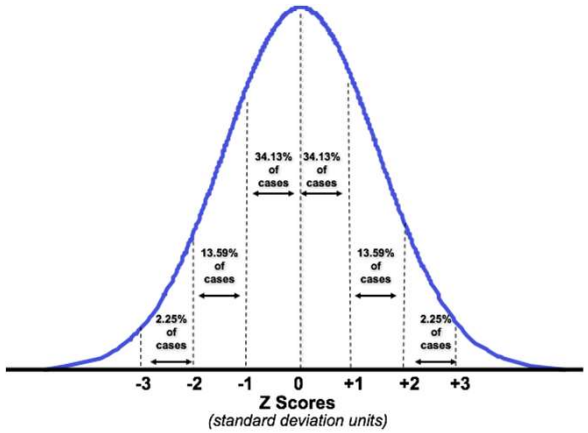
Welcome to the statistics module series

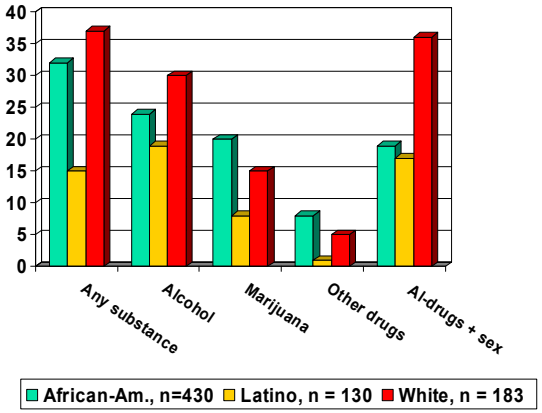
This series has seven modules:

1. Introduction to statistics & number scales



2. The Z score and the normal distribution
3. The logic of research; Plato's Allegory of the Cave
4. Testing hypotheses: The critical ratio
5. Calculating a t score
6. Testing t: The Central Limit Theorem
7. Correlations: Measures of association





© Dr. David J. McKirnan, 2014
The University of Illinois Chicago
McKirnanUIC@gmail.com
Do not use or reproduce without permission

We will cover three areas in this module:

- Introduction to statistics
- Number scales & distributions
- Central Tendency: **Mode, Median, Mean**

Numbers in science

- Number scales & distributions
- Central Tendency: **Mode, Median, Mean**

Why are numbers & statistics so important?

Science tries to explain events in the natural world.

Shutterstock.com

Events in the natural world.

Why are numbers & statistics so important?

Events in the natural world.

Operational Definitions

Numerical description of reality.

To study natural events we **operationally define** them by representing them as numbers.

This can take many forms…

- Physical measurements
- Bio-medical data; Blood pressure
- Attitude rating scales
- Counts of observed behavior

Why are numbers & statistics so important?

Numerical description of reality.

Numerical Descriptions:

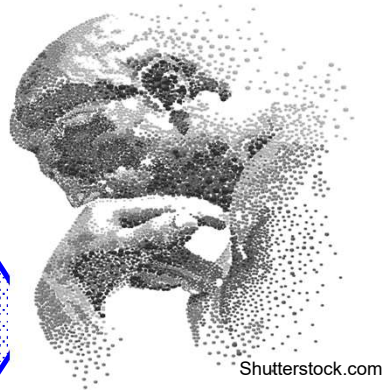
- ✓ The simple frequency of a key event;
- ✓ The central tendency – e.g., average – of a group of observations;
- ✓ The range or amount of variance in a group of observations....

Why numbers and statistics?

- ✓ How large is, say, the difference between my experimental and control groups?
- ✓ How confident can I be that my results are "real" rather than due to simple chance?
- ✓ How well can I generalize from my sample to a larger population?
- ✓ What other important research questions do these results lead to?

Numerical description of reality.

Statistics



Shutterstock.com

- Decisions.
- Conclusions.
- Further research questions.

Statistics allow us to organize and resent numbers to reach conclusions about them.

What does statistics do?

Events in the natural world.

Numerical description of reality.

- **Decisions.**
- **Conclusions.**
- **Further research questions.**

Characterizing our world is most efficient by numerical representation and analyses.

Of course numerical / statistical portrayals of the world have both advantages and disadvantages.

Costs and & benefits of numbers

Virtues of numerical data:

- **Using numbers to operationally define natural events helps:**
  - **Make our thoughts more clear and specific,**
  - **Simplify communication.**
- **Simple statistics can show the general trend or "common denominator" in a set of observations.**
- **Statistics allow us to formally test hypotheses.**
- **Statistics are hot!**
  - **According to Paul Rudman** *"statisticians are the new sexy vampires, only even more pasty…"*

    (Rudman is writing about the ability of Nate Silver, a statistician, in almost perfectly predicting the outcome of elections.)
  - **Business Insider describes 5 cool statistical problems you need to know.** (Click "show on single page" to see the article.)

Virtues of numerical data

Dangers of numerical data:

- **Our data may be affected by biases we are not aware of.**
  - We may over or under-sample certain social groups out of convenience
  - People willing to participate in research at all may be non-representative.
- **We choose which number(s) we use to describe the world.**
  - Single numbers may lead us to over-simplify the world:
  - Our choice of one number rather than another may reflect a personal or ideological bias,
- **Possible bias in the interpretation of ambiguous data:**
  - The interpretation of measurement or correlational data is prey to personal bias or expectations
- **Theory & hypotheses are more important than numbers**
  - Having a clear hypothesis helps us differentiate junk from real findings…
  - Otherwise any number pattern makes as much sense as any other.

Dangers of numerical data:

- **Our data may be affected by unintentional biases.**
  - **We may over / under-sample certain groups out of convenience**
    - **Student samples can be very biased, toward…**
      - ✓ …middle & upper-middle class, white young people,
      - ✓ …who are typically unmarried, more liberal, not “established”.

Click for a Salon Article on the resegregation of American schools

…Los Angeles [has] … the nation's largest population of chronically homeless people, a federal study has found.
(Irfan Khan / Los Angeles Times)

    - **Even the census cannot reach every American, and may have systematic biases:**
      - ✓ Transient, homeless or very poor people are hard to reach;
      - ✓ The very wealthy typically are often difficult to reach.

Dangers of numerical data:

- **Our data may be affected by unintentional biases.**
  - **We may over / under-sample certain groups out of convenience**
    - **Student samples can be very biased.**
    - **Even the census cannot reach every American, and may have systematic biases:**
  - **People willing to participate in research at all may be non-representative.**
    - **Many recruitment methods bias samples toward poorer people**
      - ✓ Using $, health care, or social services as incentives pulls for less affluent participants in many studies;
      - ✓ Professional 'Guinea Pigs' can make a living from participating in multiple clinical trials (Click image for NPR report).
    - **"Volunteerism" may bias results toward the more socially involved.**

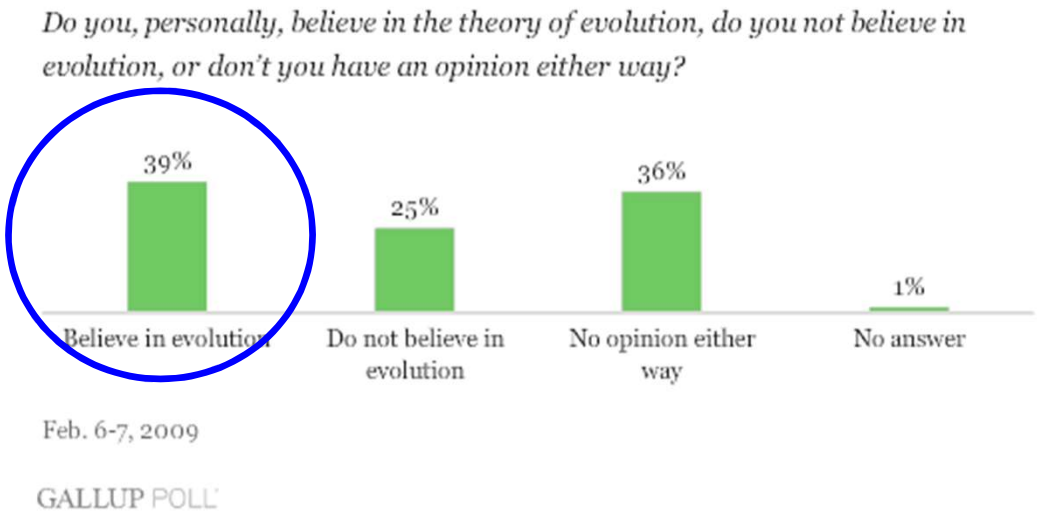
Click left image for overview.

Putting experimental drugs to the test can be a way of life.
Glow Wellness/Glow RM/Getty Images

Dangers of numerical data:

- Our data may be affected by biases we are not aware of.
- **We choose which number(s) we use to describe the world.**
  - **Single numbers may lead us to over simplify the world:**
    - We may ignore important sub-group differences:

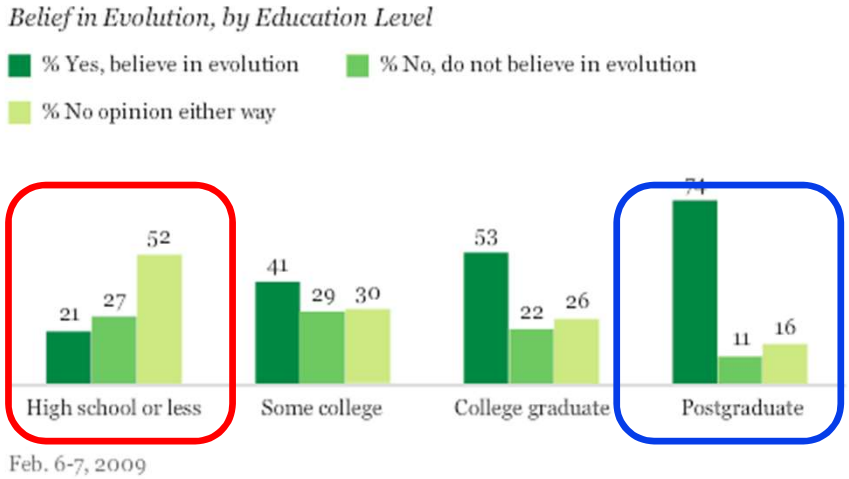
Single numbers may lead us to over-simplify the world:


Do you, personally, believe in the theory of evolution, do you not believe in evolution, or don't you have an opinion either way?
39%
25%
36%
1%
Believe in evolution
Do not believe in evolution
No opinion either way
No answer
Feb. 6-7, 2009
GALLUP POLL

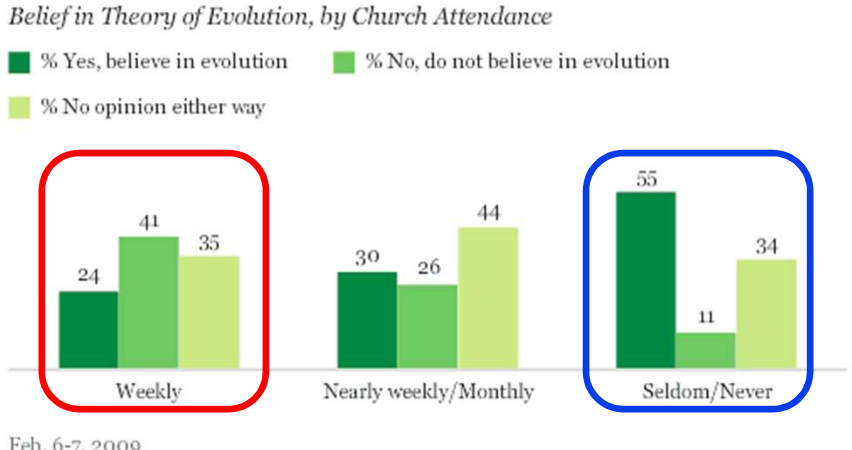
A simple question about the % of Americans who believe in evolution…

…masks important sub-group differences by:

- ✓ **Education level**
- ✓ **Church attendance**

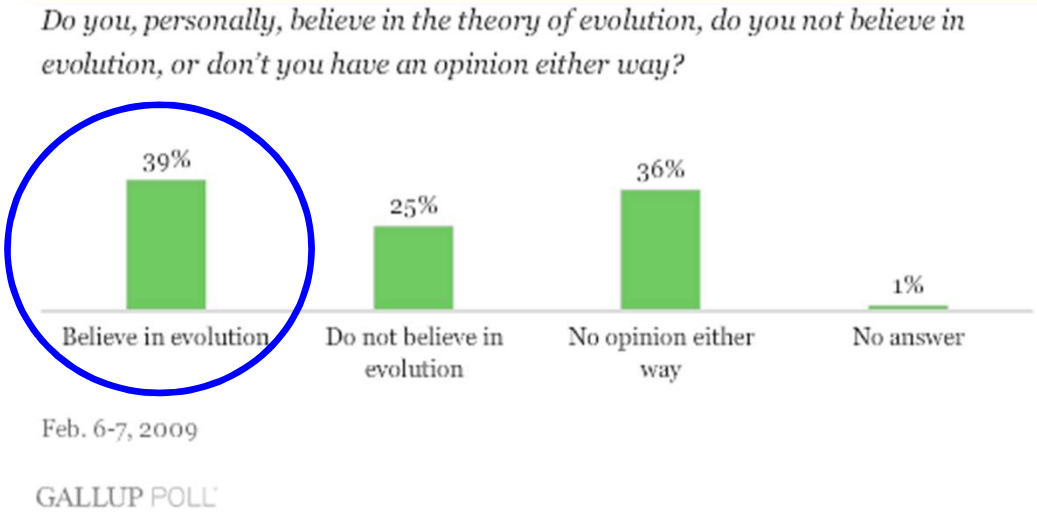


More educated groups are much more accepting of evolution.

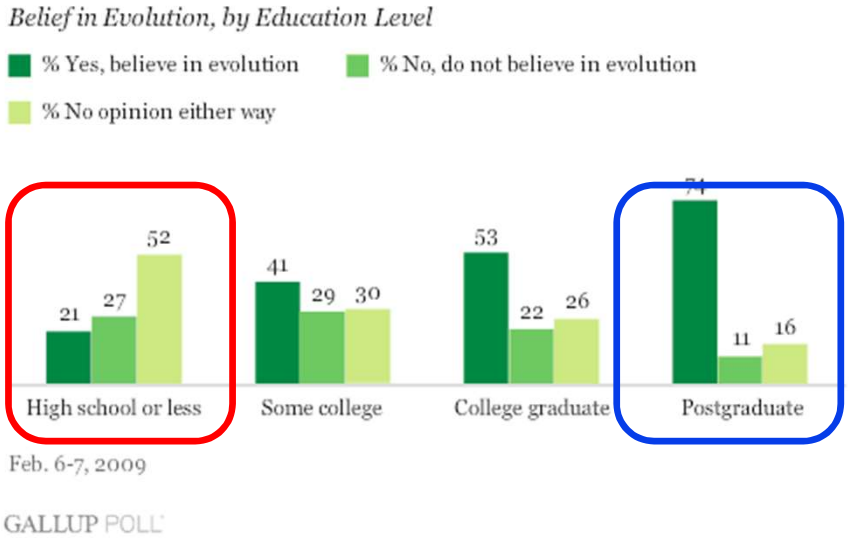


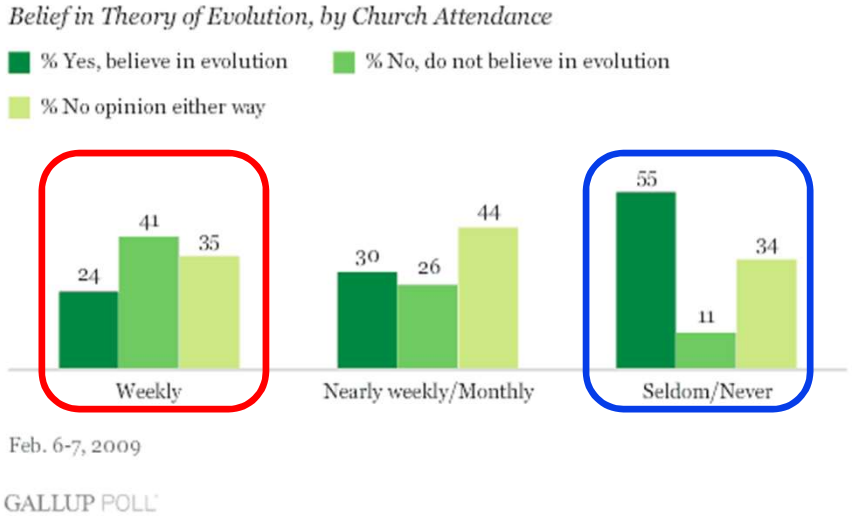
More religious groups are much less accepting of evolution.

…over-simplify or ignore sub-group differences;

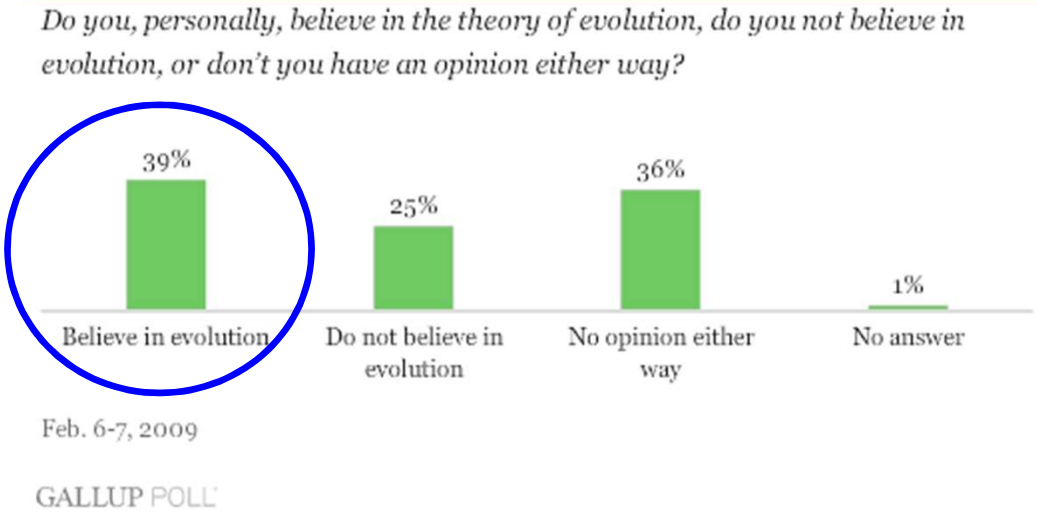

Do you, personally, believe in the theory of evolution, do you not believe in evolution, or don't you have an opinion either way?
39%
25%
36%
1%
Believe in evolution
Do not believe in evolution
No opinion either way
No answer
Feb. 6-7, 2009
GALLUP POLL

Since the overall % ignores these and other social group differences, it may not accurately represent any specific group in the sample.



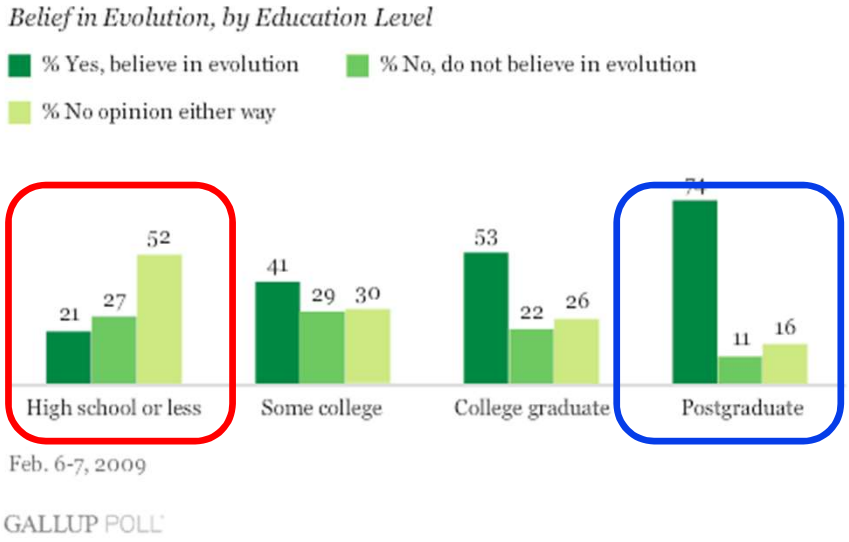


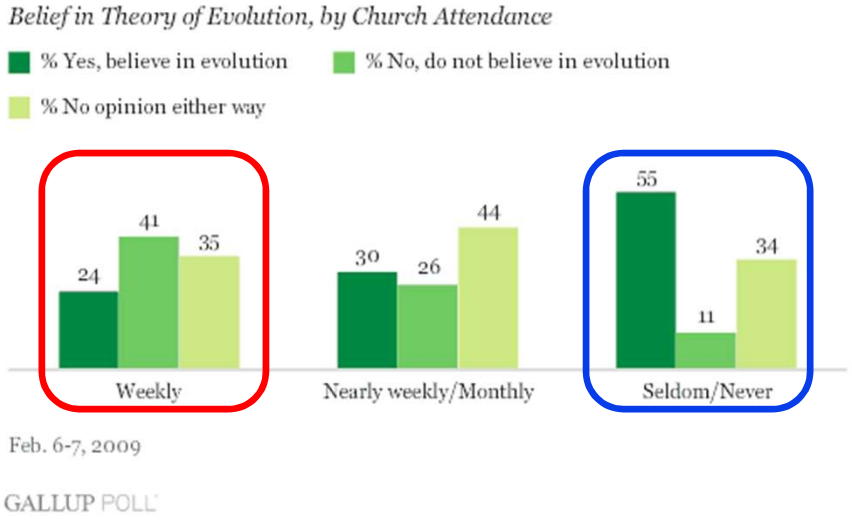
…over-simplify or ignore sub-group differences;


Do you, personally, believe in the theory of evolution, do you not believe in evolution, or don't you have an opinion either way?
39%
25%
36%
1%
Believe in evolution
Do not believe in evolution
No opinion either way
No answer
Feb. 6-7, 2009
GALLUP POLL

Using a single number to describe a complex phenomenon may be technically ("statistically") accurate…

…but incomplete.





Dangers of numerical data:

- Our data may be affected by biases we are not aware of.
- **We choose which number(s) we use to describe the world.**
  - **Single numbers may lead us to over simplify the world:**
    - Ignore important sub-group differences.
    - **…over-simplify a complex phenomenon.**

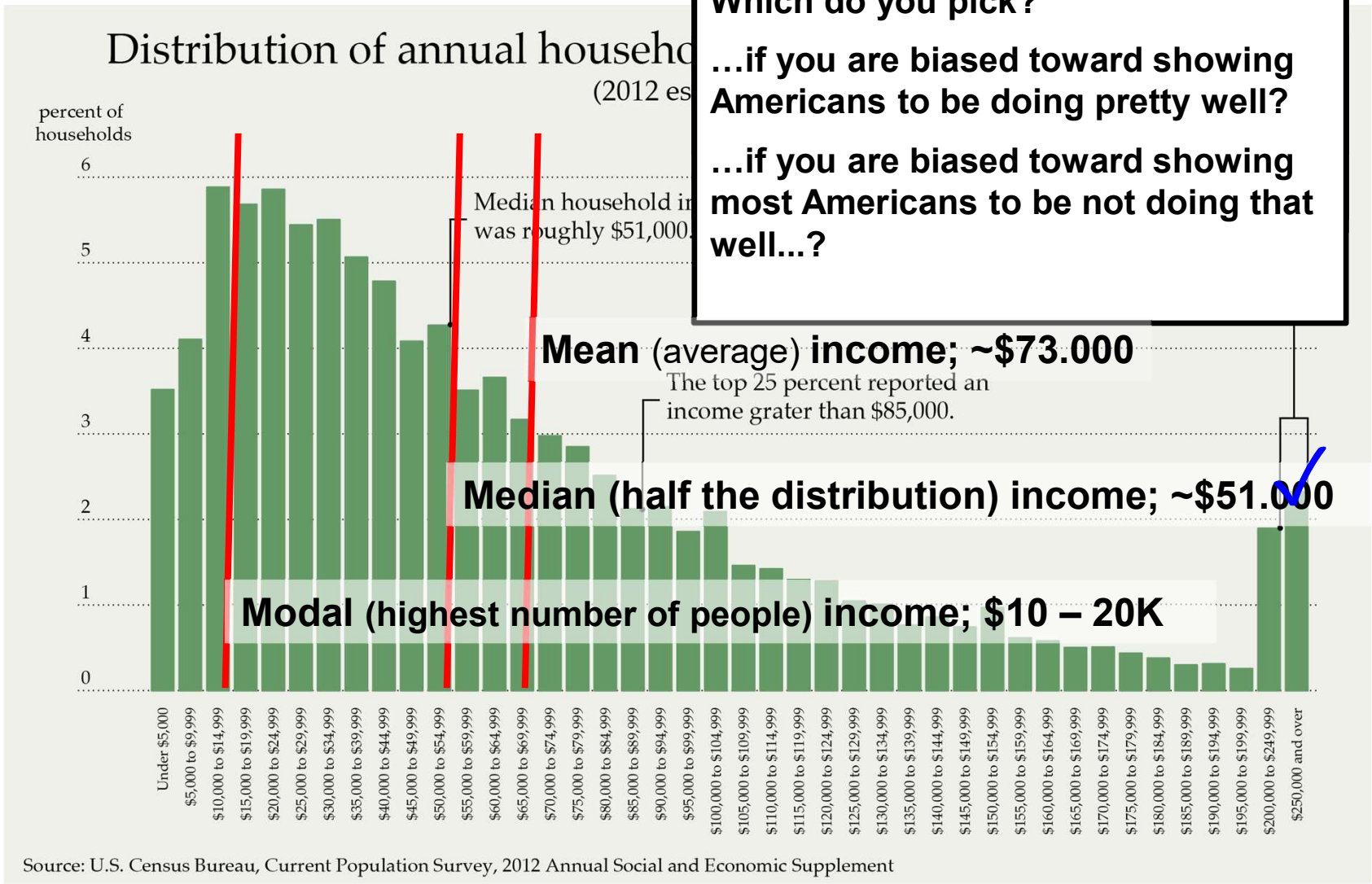
Simplifying a complex issue: How much do Americans make?

Income is highly skewed; Characterizing American incomes critically depends on which number we use.

Which do you pick?

…if you are biased toward showing Americans to be doing pretty well?

…if you are biased toward showing most Americans to be not doing that well…?



Source: U.S. Census Bureau, Current Population Survey, 2012 Annual Social and Economic Supplement

Dangers of numerical data:

- Our data may be affected by biases we are not aware of.
- **We choose which number(s) we use to describe the world.**
  - Single numbers may lead us to over simplify the world.
  - **Our choice of one number rather than another may reflect a personal or ideological bias.**
    - Are the data a biased search for ("Cherry picking") results that confirm a hypothesis?

Is the Affordable Care Act working?

…a legitimate question.

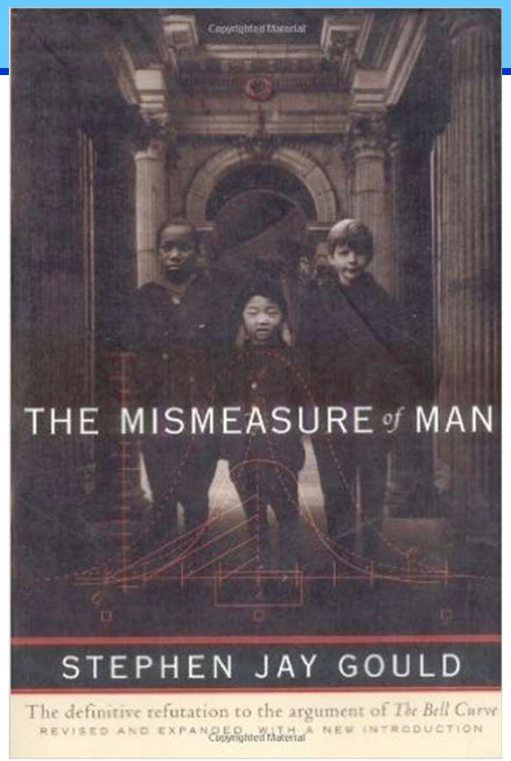
- **No; insurance rates are rising quickly.**
- **Yes; the number of uninsured has decreased substantially.**

Is climate change 'real'?

…a largely political question.

- **No; there is variance in the data and it was cold last winter.**
- **Yes; climate models show a strong effect of carbon emissions on global warming and weather extremes.**

Biased search for results that confirm a hypothesis…


THE MISMEASURE of MAN
STEPHEN JAY GOULD
The definitive refutation to the argument of The Bell Curve
REVISED AND EXPANDED, WITH A NEW INTRODUCTION

- ✓ Stephan j Gould, in *The Mismeasure of Man* describes the racist and (social) "classist" biases in the history of human assessment.
- ✓ He makes the point that ascribing individual differences – much less differences among racial groups – to heredity is a political, not a scientific perspective.
- ✓ Gould describes the commitment of IQ researchers and statisticians in the early-mid 20th century to the *a priori* conviction that intelligence is a single, unidimensional skill or ability.
- ✓ This unidimensional ability was assumed to be genetically based.
- ✓ Following this assumption, researchers created instruments and used statistical techniques such as Factor Analysis* to reduce diverse indicators of intelligence to one core dimension: General Intelligence, or the "G" factor.
- ✓ Of course this "validated" the assumption that intelligence has a single "universal" dimension, and made it easy to ignore cultural differences in the content or measurement of intelligence.

*Factor Analysis uses correlational techniques to derive the core or most common factors that underlie multiple measures of a construct.

Dangers of numerical data:

- Our data may be affected by biases we are not aware of.
- **We choose which number(s) we use to describe the world.**
  - Single numbers may lead us to over simplify the world.
  - **Our choice of one number rather than another may reflect a personal or ideological bias.**
    - Are the data a biased search for ("Cherry picking") results that confirm a hypothesis?
    - **Surveys with leading questions or high social desirability generate biased responses.**
      - "Push" polls and biased political reporting; see Surveys.

Dangers of numerical data:

- Our data may be affected by biases we are not aware of.
- **We choose which number(s) we use to describe the world.**
  - Single numbers may lead us to over simplify the world.
  - **Our choice of one number rather than another may reflect a personal or ideological bias.**
    - Are the data a biased search for ("Cherry picking") results that confirm a hypothesis?
    - Surveys with leading questions or high social desirability generate biased responses.
  - **We may simply ignore "inconvenient" data..**

Simply ignoring disconfirming or "inconvenient" data:

Project DARE continues as a politically favored but ineffective drug abuse prevention program. (Click for a review)

- ✓ Dare maintains political support as a criminal justice –based program.
- ✓ Data showing the program to have little or no effect on drug use is often discounted by pointing to other potentially positive effects, e.g., on the community.

Data indicating climate change are ignored or distorted by many political commentators.

- ✓ The U.S. is distinctive in the industrialized world in framing climate change as a political "wedge" issue.

Relax: Global warming is all a myth was The Spectator's cover for a story on Ian Plimer. Photograph: The Spectator

Downloaded from http://www.theguardian.com/environment/georgemonbiot/2009/sep/23/spectator-plimer-climate-change-sceptic 3/6/15

Dangers of numerical data:

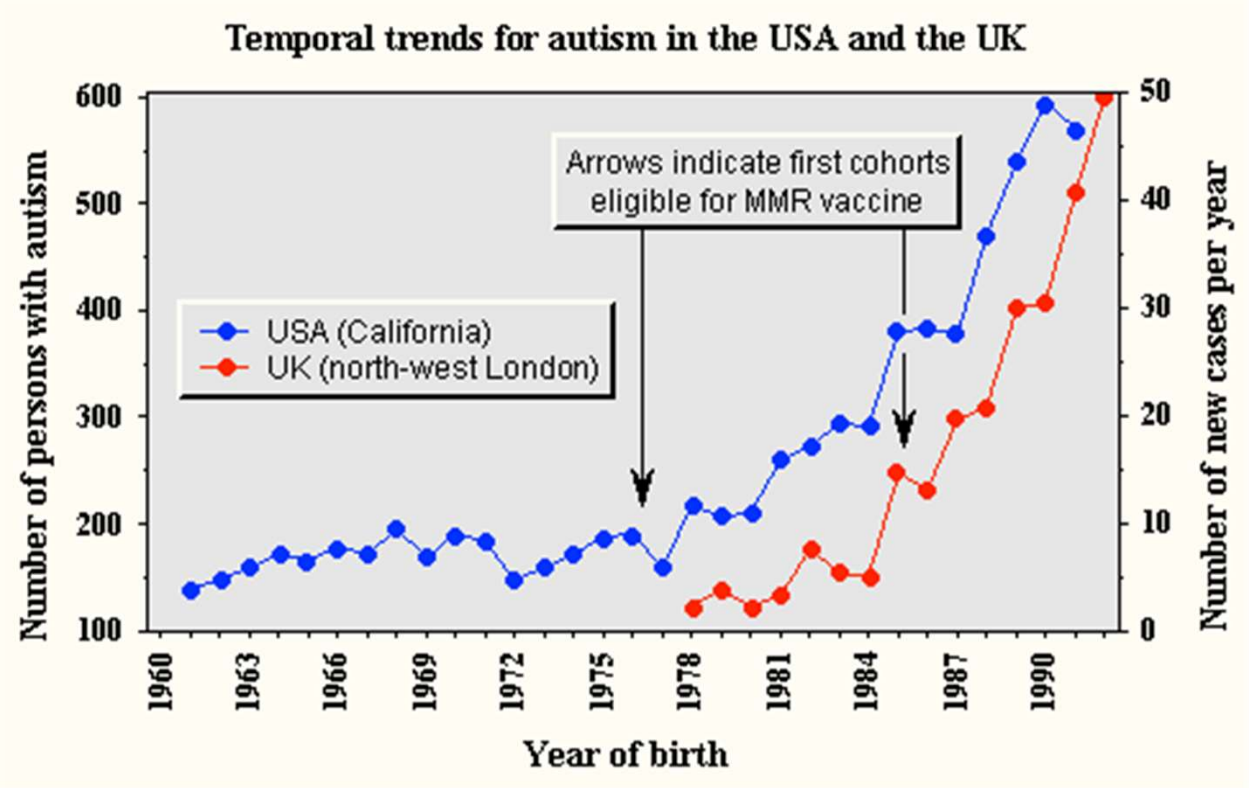
- Our data may be affected by biases we are not aware of.
- We choose which number(s) we use to describe the world.
- **Possible bias in the interpretation of ambiguous data:**
  - The interpretation of measurement or correlational data is particularly prey to personal biases or expectations.

Data patterns – e.g., crime or disease rates – can be interpreted in diverse and potentially biased ways.

Why are autism diagnoses substantially increasing over the past 20 or 30 years?

✓ **Is autism caused by the Measles, Mumps & Rubella (MMR) vaccine?**

This common graph shows autism rates in two different studies. Rates increased substantially after the MMR vaccine was introduced.



Downloaded from http://influentialpoints.com/Critiques/analysis_re_wakefield_&_mmr.htm 3/6/15.

This data pattern appears to show the MMR vaccine "causes" autism; rates go up after the vaccine is introduced.

Political (mis)uses of scientific data

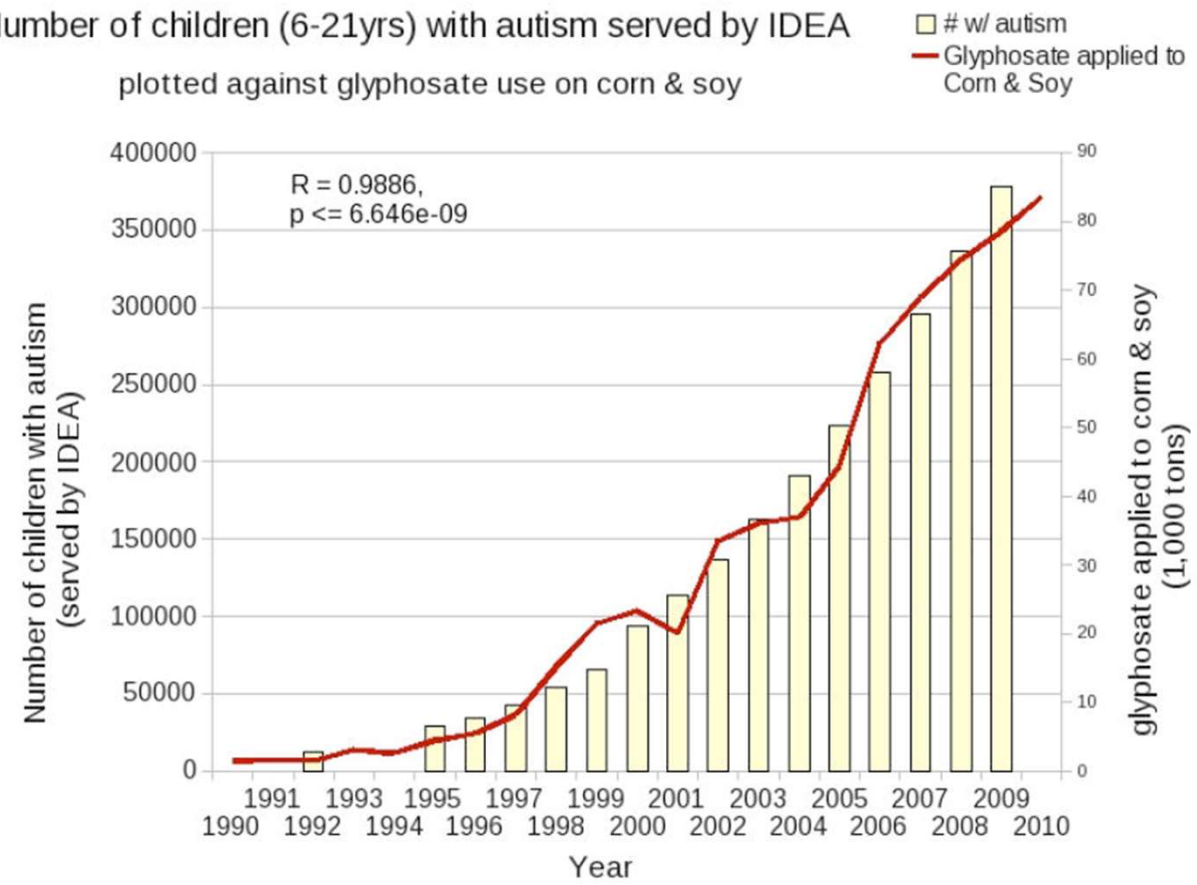
Data patterns – e.g., crime or disease rates – can be interpreted in diverse and potentially biased ways.

Why are autism diagnoses substantially increasing over the past 20 or 30 years?

- ✓ Is autism caused by the Measles, Mumps & Rubella (MMR) vaccine?
- ✓ **Perhaps autism is caused by the increased use of glyphosphate, an agricultural herbicide…**

"Happenstance" correlations; Autism and glyphosphate

Could herbicide use (often associated with genetically modified crops) **cause disease?**



Downloaded from; http://geneticliteracyproject.org/2015/01/autism-increase-mystery-solved-no-its-not-vaccines-gmos-glyphosate-or-organic-foods/ 3/6/15. See: http://farmwars.info/?p=11700.

Here the use of glyphosphate herbicide tracks closely with the increase in autism diagnosis.

Does one cause the other?

Political (mis)uses of scientific data

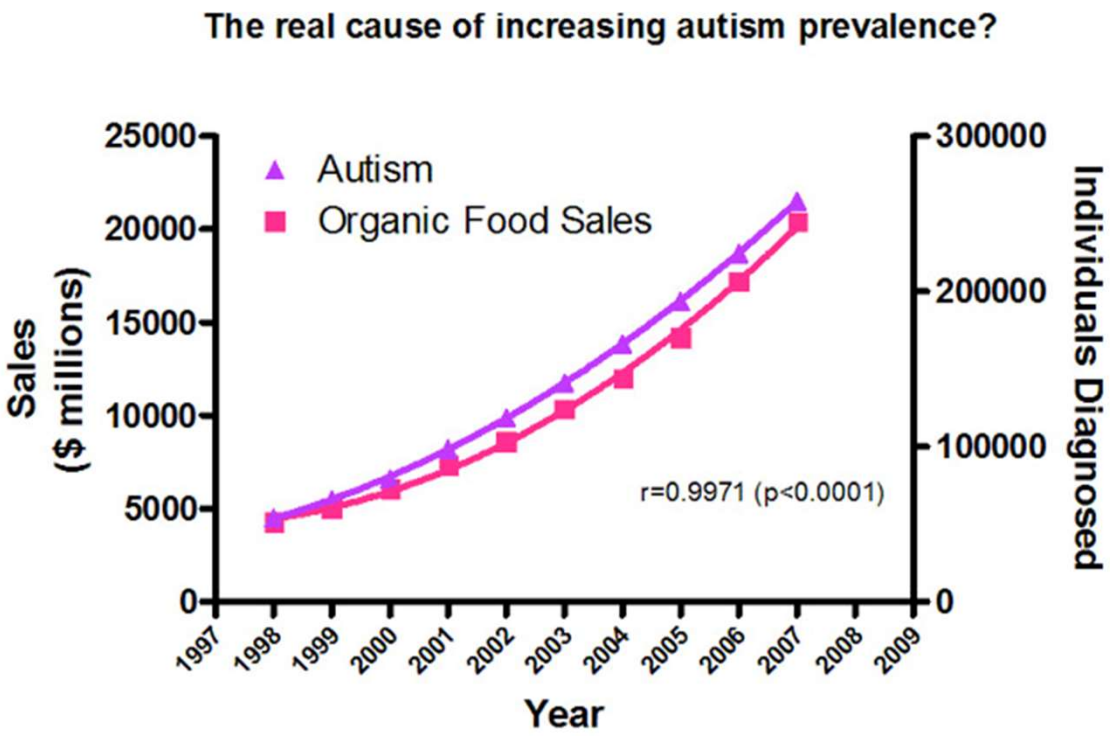
Data patterns – e.g., crime or disease rates – can be interpreted in diverse and potentially biased ways.

Why are autism diagnoses substantially increasing over the past 20 or 30 years?

- ✓ Is autism caused by the Measles, Mumps & Rubella (MMR) vaccine?
- ✓ Is autism caused by the increased use of glyphosphate, an agricultural herbicide?
- ✓ **Perhaps autism is caused by eating organic food…**

"Happenstance" correlations; Autism and organic food

Perhaps autism is caused by eating organic food…



Downloaded from: http://geneticliteracyproject.org/2015/01/autism-increase-mystery-solved-no-its-not-vaccines-gmos-glyphosate-or-organic-foods/ 3/6/15

Autism diagnoses happen to correlate with organic food sales.

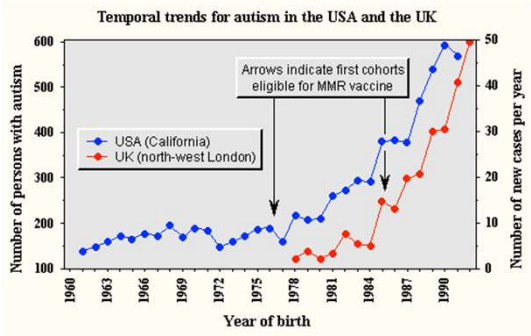
- ✓ **We intuitively want to figure out how organic foods may cause autism.**
- ✓ **Causality may be irrelevant here:**
- ✓ **They both are simply increasing over the same time period…**

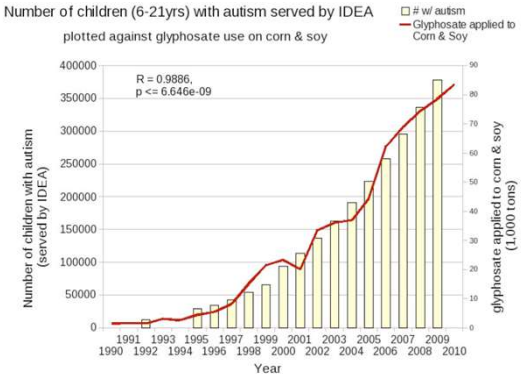
Political (mis)uses of scientific data

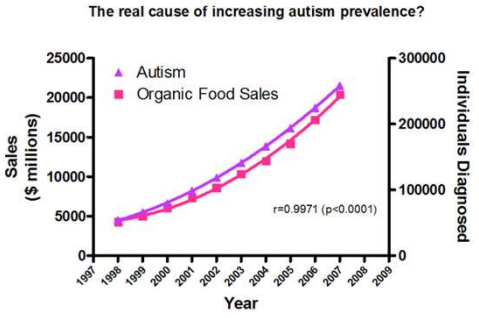
Data patterns – e.g., crime or disease rates – can be interpreted in diverse and potentially biased ways.

Why are autism diagnoses substantially increasing over the past 20 or 30 years?

✓ **All these interpretations take ambiguous correlations and apply a political or ideological agenda to them.**

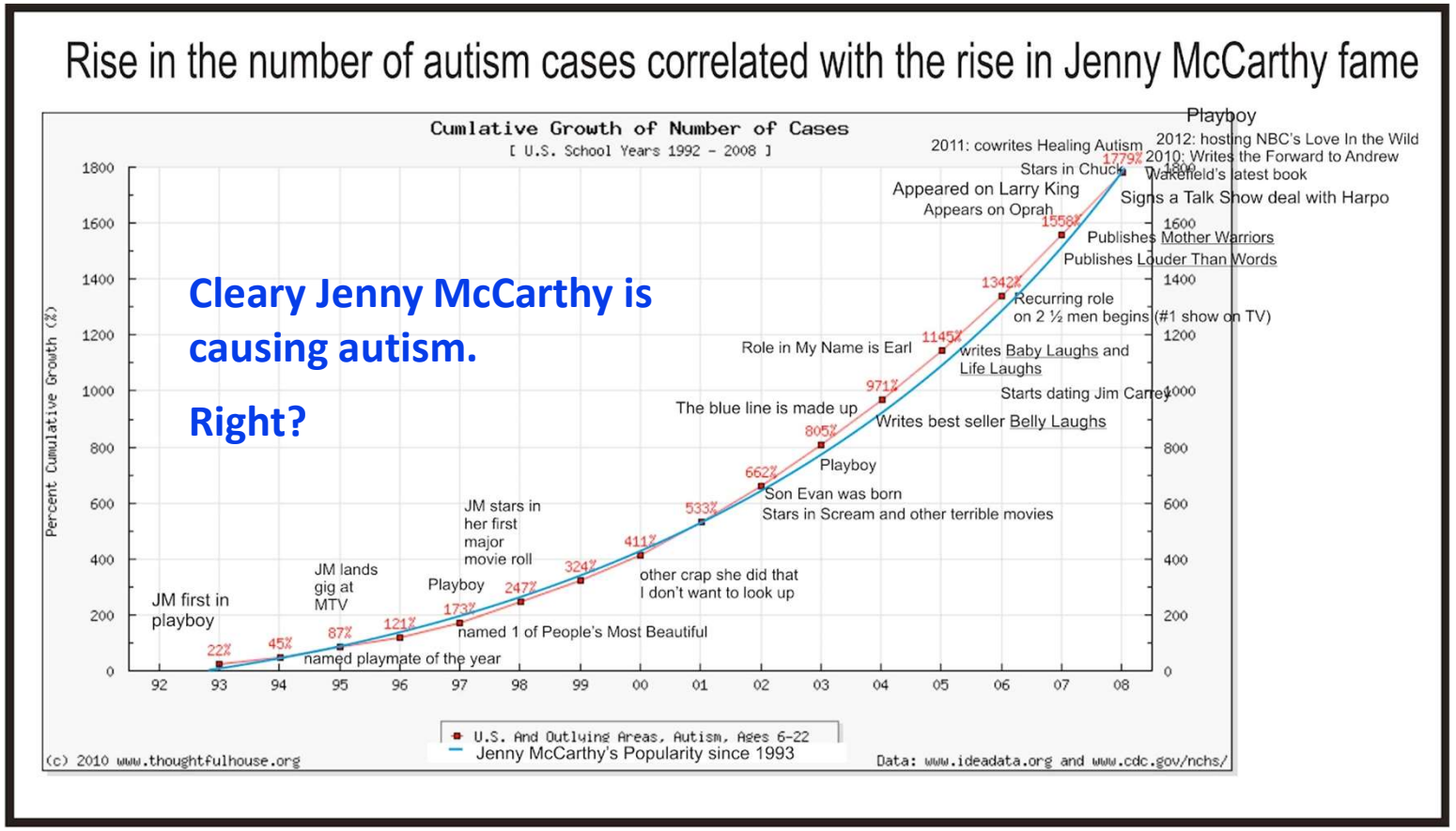

Temporal trends for autism in the USA and the UK
Arrows indicate first cohorts eligible for MMR vaccine
USA (California)
UK (north-west London)
Number of persons with autism
Number of new cases per year
Year of birth


Number of children (6-21yrs) with autism served by IDEA
plotted against glyphosate use on corn & soy
w/ autism
Glyphosate applied to Corn & Soy
R = 0.9886,
p <= 6.646e-09
Number of children with autism (served by IDEA)
glyphosate applied to corn & soy (1,000 tons)
Year



Interpretations of ambiguous correlations

Autism incidence even strongly correlates with the fame of Jenny McCarthy* (a celebrity anti-vaccine advocate)**.**


Rise in the number of autism cases correlated with the rise in Jenny McCarthy fame
Cumlative Growth of Number of Cases
[U.S. School Years 1992 - 2008]
Percent Cumulative Growth (%)
Cleary Jenny McCarthy is causing autism.
Right?
JM first in playboy
JM lands gig at MTV
named playmate of the year
Playboy
JM stars in her first major movie roll
named 1 of People's Most Beautiful
other crap she did that I don't want to look up
Son Evan was born
Stars in Scream and other terrible movies
Playboy
The blue line is made up
Writes best seller Belly Laughs
Role in My Name is Earl
writes Baby Laughs and Life Laughs
Starts dating Jim Carrey
Recurring role on 2 ½ men begins (#1 show on TV)
Publishes Mother Warriors
Publishes Louder Than Words
Appeared on Larry King
Appears on Oprah
Stars in Chuck
2011: cowrites Healing Autism
2012: hosting NBC's Love In the Wild
2010: Writes the Forward to Andrew Wakefield's latest book
Signs a Talk Show deal with Harpo
U.S. And Outlying Areas, Autism, Ages 6-22
Jenny McCarthy's Popularity since 1993
(c) 2010 www.thoughtfulhouse.org
Data: www.ideadata.org and www.cdc.gov/nchs/

Downloaded from http://www.skepticblog.org/2014/01/22/autism-and-vaccines-correlation-is-not-causation/, 3/6/15

*At least in one totally made up and spurious way of tracking her fame…

Political (mis)uses of scientific data

Data patterns – e.g., crime or disease rates – can be interpreted in diverse and potentially biased ways.

Why are autism diagnoses substantially increasing over the past 20 or 30 years?

In fact, the criteria for the diagnosis of "autism spectrum disorder" have expanded several times in the past 20 years.

- ✓ *The evidence … suggests that most, if not all, of the reported increase … is due to changes in diagnostic criteria and greater awareness among both parents and professionals.*
- ✓ *It remains an open question whether there has been a genuine rise in the numbers of children of autistic spectrum disorders…*

Wing, L. & Potter, D. (2002), THE Epidemiology of Autistic Spectrum Disorders: Is The Prevalence Rising? MENTAL RETARDATION AND DEVELOPMENTAL DISABILITIES RESEARCH REVIEWS 8: 151–161. P. 159

Dangers of numerical data:

- Our data may be affected by biases we are not aware of.
- We choose which number(s) we use to describe the world.
- Possible bias in the interpretation of ambiguous data.
- **Theory & hypotheses are more important than numbers.**
  - Having a clear hypothesis helps us differentiate junk from real findings…
  - Otherwise any number pattern makes as much sense as any other.

Values:
Critical thought + Empiricism
- ✓ Understand the Natural World
- ✓ Theory: How or Why?
- ✓ Evidence: How do you know?

The primary goal of science:
"How does nature 'work"?

Numbers are a tool, not an end.

Content
- ✓ Empirical findings: Facts
- ✓ Ways of classifying nature
- ✓ Well supported theories
- ✓ Science is public

Methods
- ✓ Objective approaches
- ✓ Basic experimental design
- ✓ Specific research procedures

Dangers of numerical data:

- Our data may be affected by biases we are not aware of.
- We choose which number(s) we use to describe the world.
- Possible bias in the interpretation of ambiguous data.
- Theory & hypotheses are more important than numbers.

Thus…

- **Representing nature in terms of numbers is important to scientific progress**
  - Key for operational definitions
  - Provide powerful analytic & communication tools
- **Understanding ANY statistic requires that we understand how it was derived**
  - Statistical stateme… misleading.
    - What context w…
    - What else do we need to know to evaluate it?
    - How do we interpret ambiguous correlations?
  - Political or commercial misuse of numerical / statistical information is common and problematic.

> **There are lies, damned lies, and statistics.**
>
> -- Mark Twain

- **Variable** → **Characteristic or attribute with different levels or qualities (e.g., age, speed, effort, attention).**

Participant	Age
Jo	32
Sally	45
Bill	22
Sam	37
Mary Louise…	34…

Some basic terms:

- Variable → Characteristic or attribute with different levels or qualities (e.g., age, speed, effort, attention).
- **Value** → **One level or state of a variable (e.g., age = 21, ethnicity = Latino).**

Participant	Age
Jo	32
Sally	45
Bill	22
Sam	37
Mary Louise…	34…

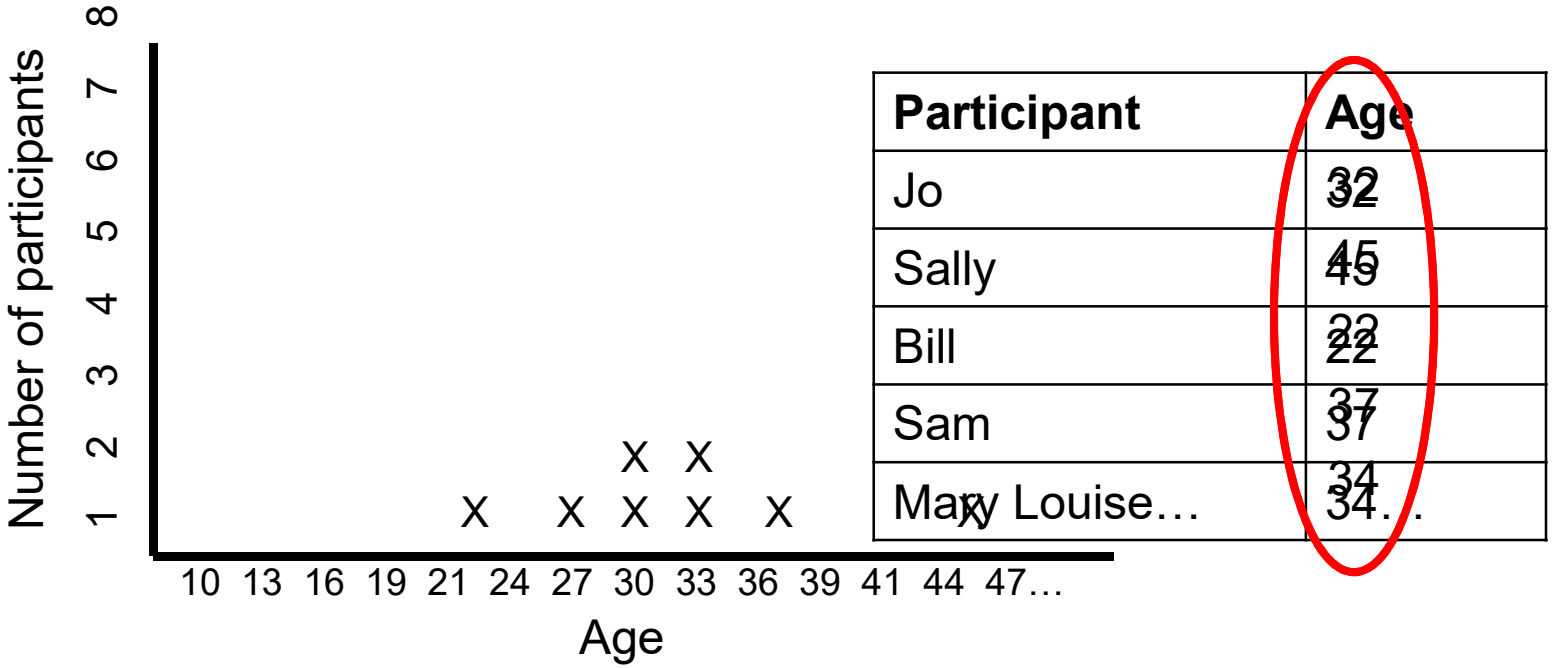
Some basic terms:

- **Variable** → Characteristic or attribute with different levels or qualities (e.g., age, speed, effort, attention).
- **Value** → One level or state of a variable (e.g., age = 21, ethnicity = Latino).
- **Distribution** → **Set of scores (each with its own *value*) for one *variable* (e.g., distribution of student ages).**

Participant	Age
Jo	32
Sally	45
Bill	22
Sam	37
Mary Louise…	34…

Some basic terms:

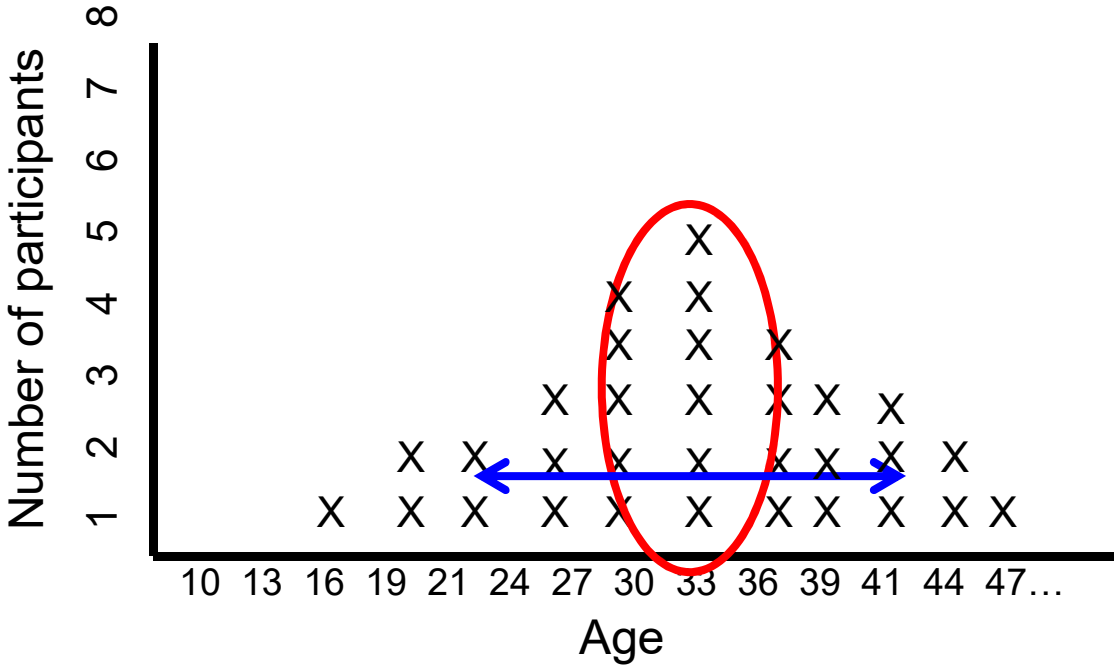
- **Variable** → Characteristic or attribute with different levels or qualities (e.g., age, speed, effort, attention).
- **Value** → One level or state of a variable (e.g., age = 21, ethnicity = Latino).
- **Distribution** → **Set of scores (each with its own *value*) for one *variable* (e.g., distribution of student ages).**



Participant	Age
Jo	32
Sally	45
Bill	22
Sam	37
Mary Louise…	34…

Some basic terms:

- **Central Tendency** → **Primary "drift" of a set of scores (*Mean*, *Mode*, or *Median*).**
- **Variance** → **Measure of how much the scores in a distribution differ from each other.**



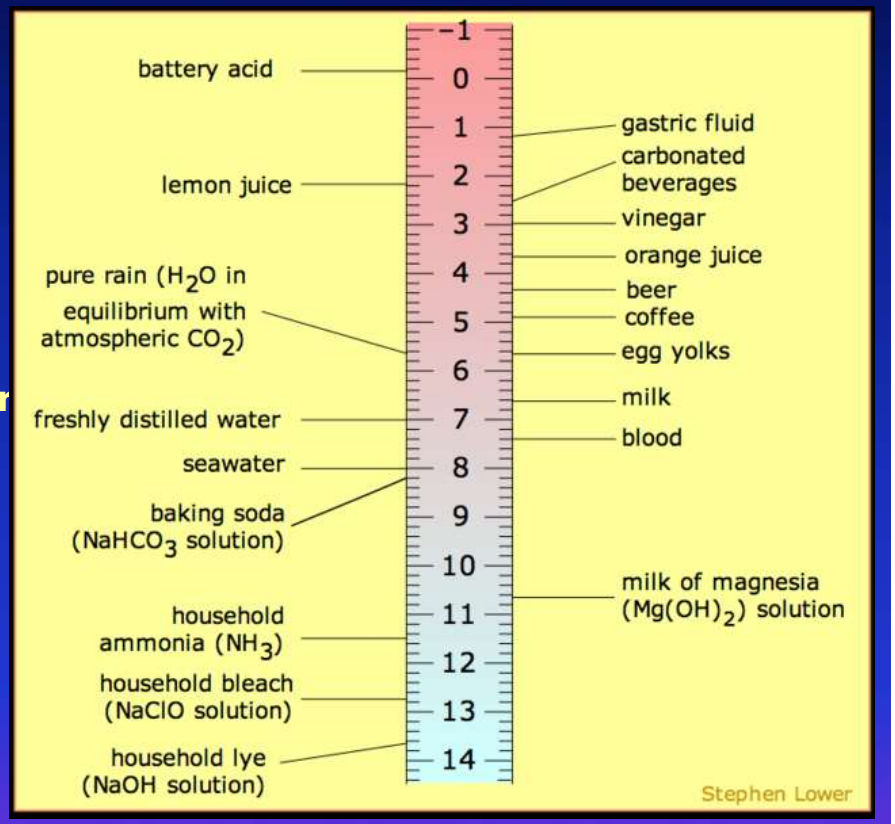
Some basic terms:

- **Variable** → Characteristic or attribute with different levels or qualities (e.g., age, speed, effort, attention).
- **Value** → One level or state of a variable (e.g., age = 21, ethnicity = Latino).
- **Distribution** → Set of scores (each with its own *value*) for one *variable* (e.g., distribution of student ages).
- **Central Tendency** → Primary "drift" of a set of scores (*Mean*, *Mode*, or *Median*).
- **Variance** → Measure of how much the scores in a distribution differ from each other; overall dispersion of scores.
- **Parameter** → **Mathematical characteristic of (a variable that describes an aspect of…) a population.**
- **Statistic** → **Mathematical characteristic of a sample.**

- Using numbers in science

☑ Number scales & frequency distributions

- Central Tendency: **Mode, Median, Mean**
- Variance: **Standard Deviation**
- The **Z score** and the normal distribution
- Using Z scores to evaluate data
- Testing hypotheses: **critical ratio**.



Numerical **scales used in research**

Scientific research uses a variety of numerical scales to assess or operationally define phenomena.

We will overview the four basic scale types here:

More advanced scales are typically based on these.

For example, we use to a Logarithmic transform to capture very large quantities, expressing them on a $\log_{10}$ scale**.**

Ratio
Interval
Ordinal
Categorical

Numerical scales

Ratio

Scales of physical properties.

Used for physical description: temperature, elapsed time, height

Interval

Arbitrary or relative ψ scales

Common in behavioral research, e.g., attitude or rating scales.

Ordinal

Rank order with non-equal intervals

Simple finish place, rank in organization, most, 2nd most, 3rd most...

Categorical

Categories only

Typical of inherent categories: ethnic group, gender, zip code

Continuous Scales

Types of numerical scales

Ratio

Scales of physical properties.

Interval

Arbitrary or relative ψ scales

Ordinal

Rank order with non-equal intervals

Categorical

Categories only

Continuous Scales

- ✓ There is a **range of values** (from highest to lowest);
- ✓ The numerical **values are meaningful**; 3 is more of an attribute

- ✓ Values are **arbitrary**;
- ✓ If Male = 1 and Female = 2, women do not have “more gender” than do men…

Ratio Scales

temperature, height, blood pressure..

Ratio
Interval
Ordinal
Categorical

- **Each scale value describes a physical reality**
  - **Batting average: % hits / at bats**
- **The zero point is grounded in a physical property**
  - **In baseball, 0 hits is meaningful**
  - **Right now you have 0 running speed…**

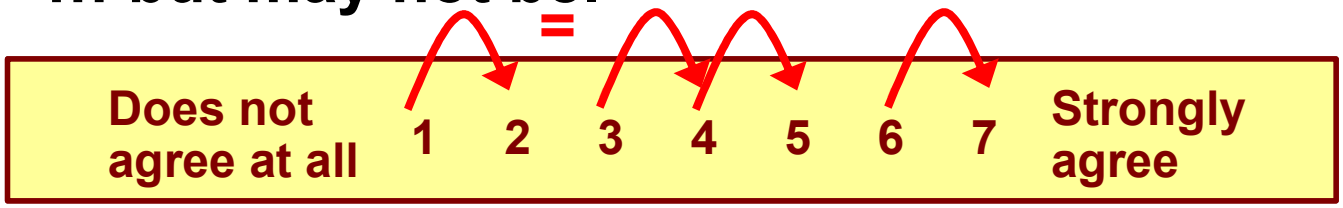
- **Each scale point is absolute** (not **relative**).
  - **Individual scale points are also grounded in physical reality.**
  - **32º has an absolute meaning, not just relative to other temperatures.**
- **Each interval is continuous & exactly equal:**
  - **30ºF → 40ºF ≡ 90ºF → 100ºF**
- **Used for correlational designs, or as dependent variable in experiments.**

Ratio
Interval
Ordinal
Categorical

Interval Scales; e.g, attitude rating scales, IQ, …

- Not grounded in physical reality; "ψ reality" only
- Correlational designs, dependent variable in experiments.
- Does **not** have an absolute "0" point; each scale point is relative
- Intervals are designed to be the same…

  … but may not be.



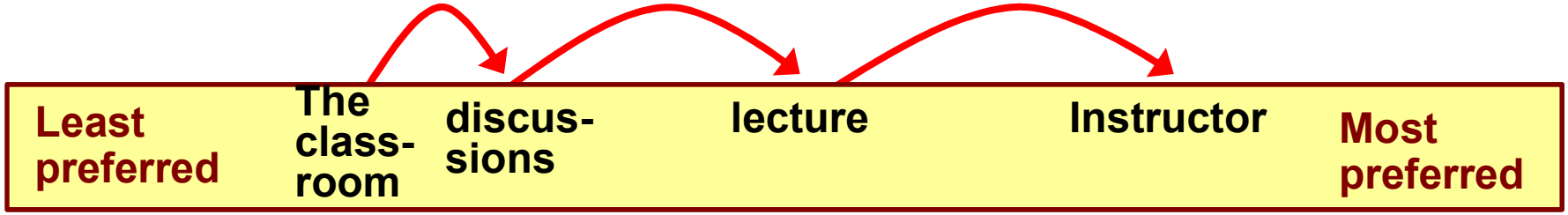
In agreeing with an attitude statement (research has changed my life…)

this is a bigger ψ step than this

Ratio
Interval
Ordinal
Categorical

Ordinal Scales; e.g., rank order

- Also not grounded in physical reality
- Relative "strength" or placement on a scale
- Primarily correlational / measurement designs.
- No "0" point; each scale point is relative
- Provides no information about intervals



Intervals between each scale point are arbitrary…

(Many interval scales may actually be ordinal…)

Ratio
Interval
Ordinal
Categorica

Categorical Scales

- **Simple groups or categories**
  - ✓ **Typical of variables in many measurement studies.**

    e.g., gender differences in epidemiology research
  - ✓ **The Independent Variable in an experiment.**

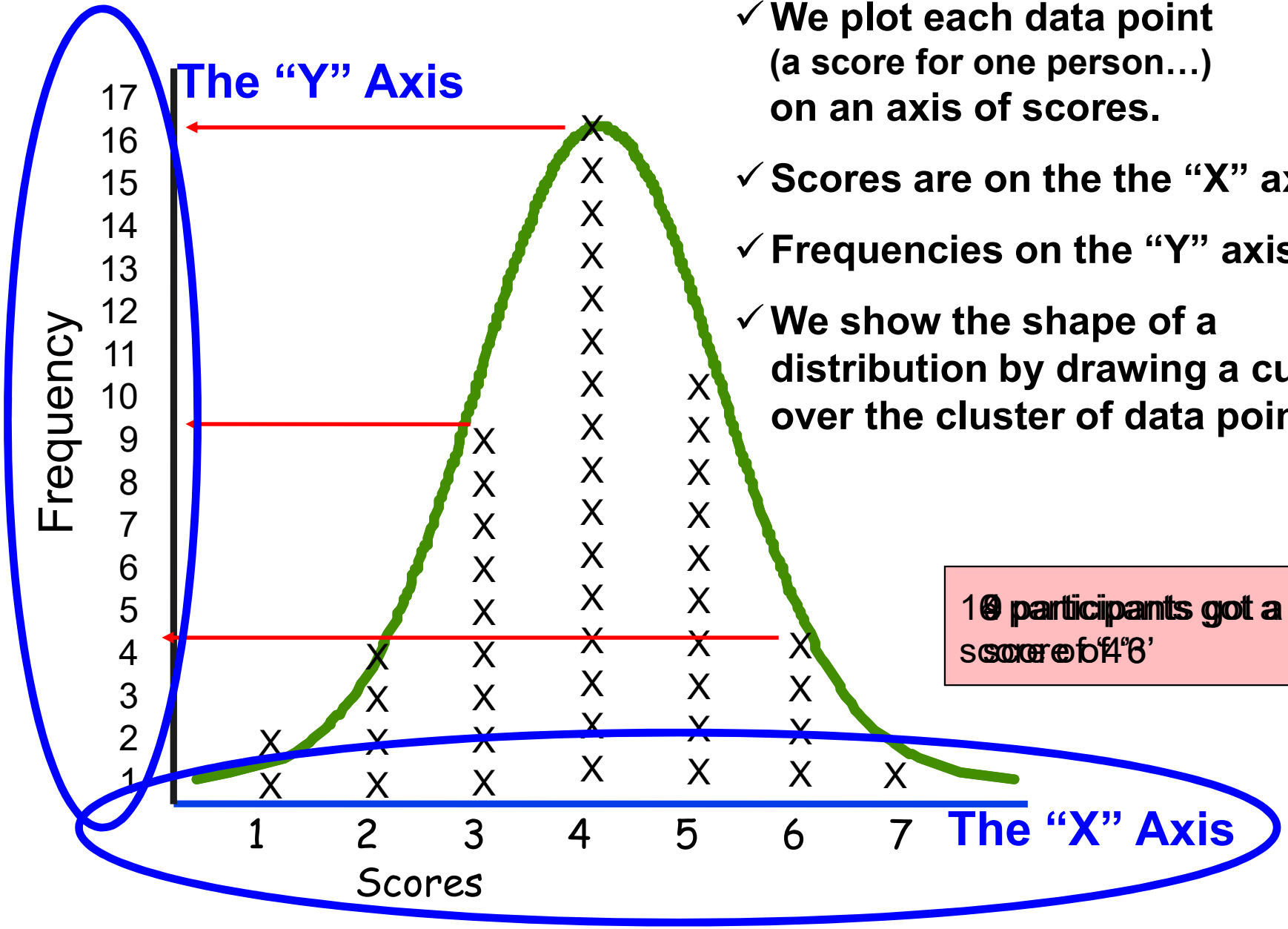
    …experimental v. control group

    …condition 1 v. condition 2 v. condition 3…
- **Binary** ("nominal") **measures: there are only 2 levels of the variable**
- **Categorical: >2 levels of the variable.**
  - **Inherent groups:** gender, ethnicity, 'handedness'…
  - **Behavioral:** neighborhood, political party, 'drug user'
  - **Arbitrary:** experimental group...

Two central ways of using numbers.

- **Descriptive Statistics:**
  - Simple quantitative description or summary.
    - Batting average in baseball
    - Grade-point average
  - **Univariate Analysis:** examines cases in terms of a single variable.
    - **Frequency distributions** group data into categories.
      - E.g., drug use by Age categories, Ethnic groups…
- **Inferential Statistics:**
  - Conduct analyses on **samples**
    - Compare groups (experimental v. control…)
    - Characterize a sample in epidemiological research
  - Use statistical operations to **generalize** the results to a **population**.

What is a Frequency Distribution?

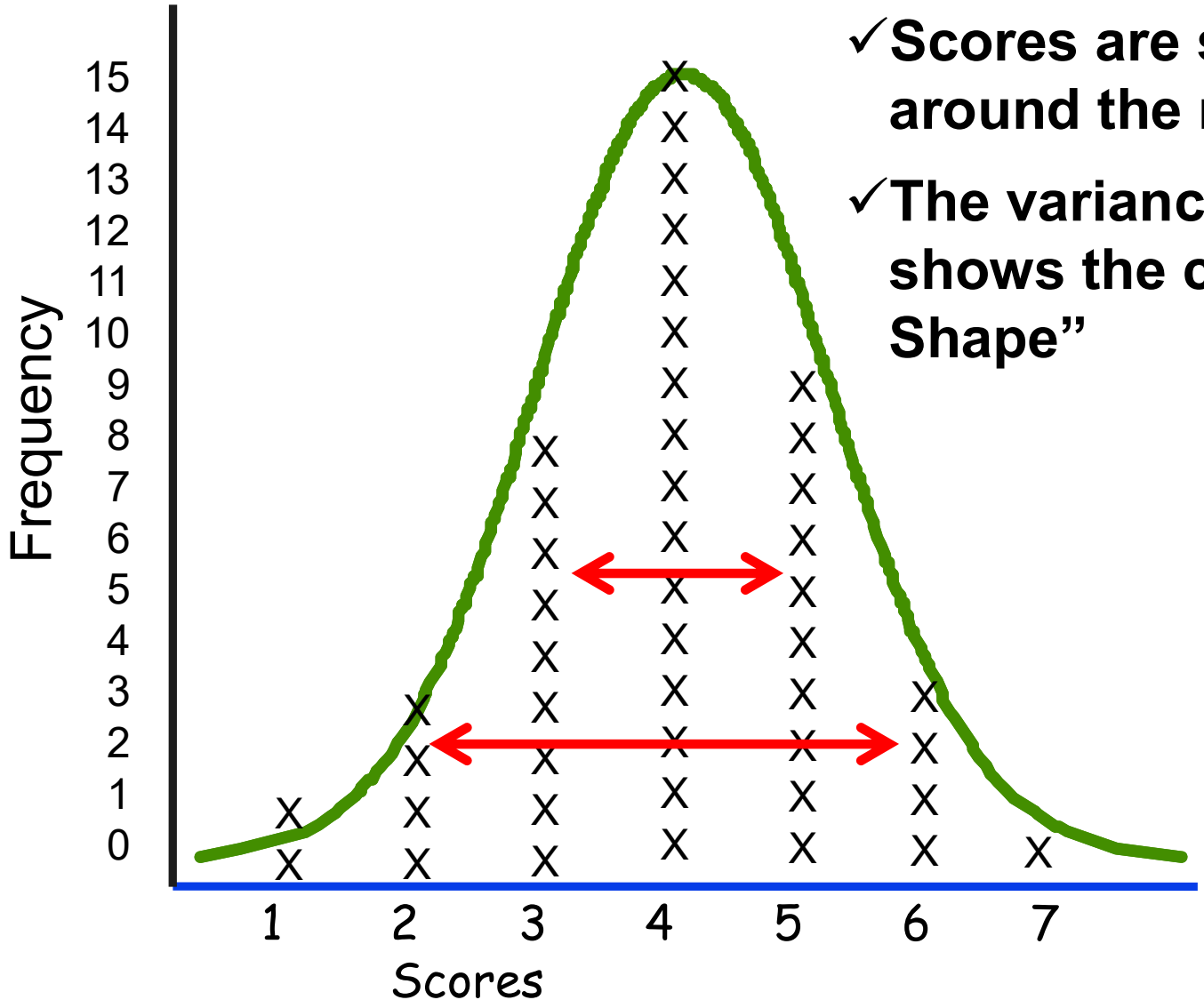


- ✓ We plot each data point (a score for one person…) on an axis of scores.
- ✓ Scores are on the the "X" axis
- ✓ Frequencies on the "Y" axis..
- ✓ We show the shape of a distribution by drawing a curve over the cluster of data points…

16 participants got a score of "4"

In a "normal" distribution

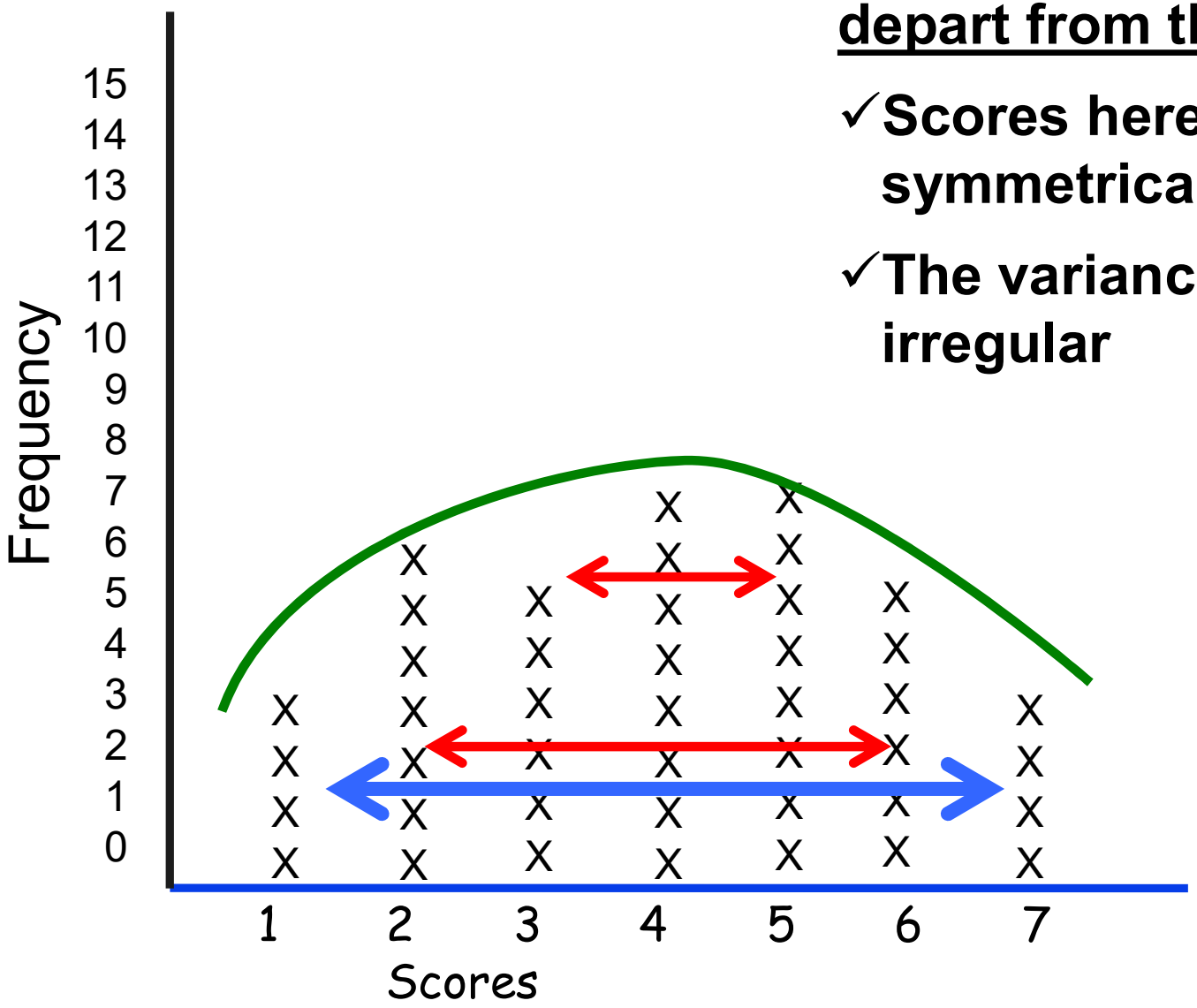
- ✓Scores are symmetrical around the mid-point
- ✓The variance in scores shows the classic "Bell Shape"

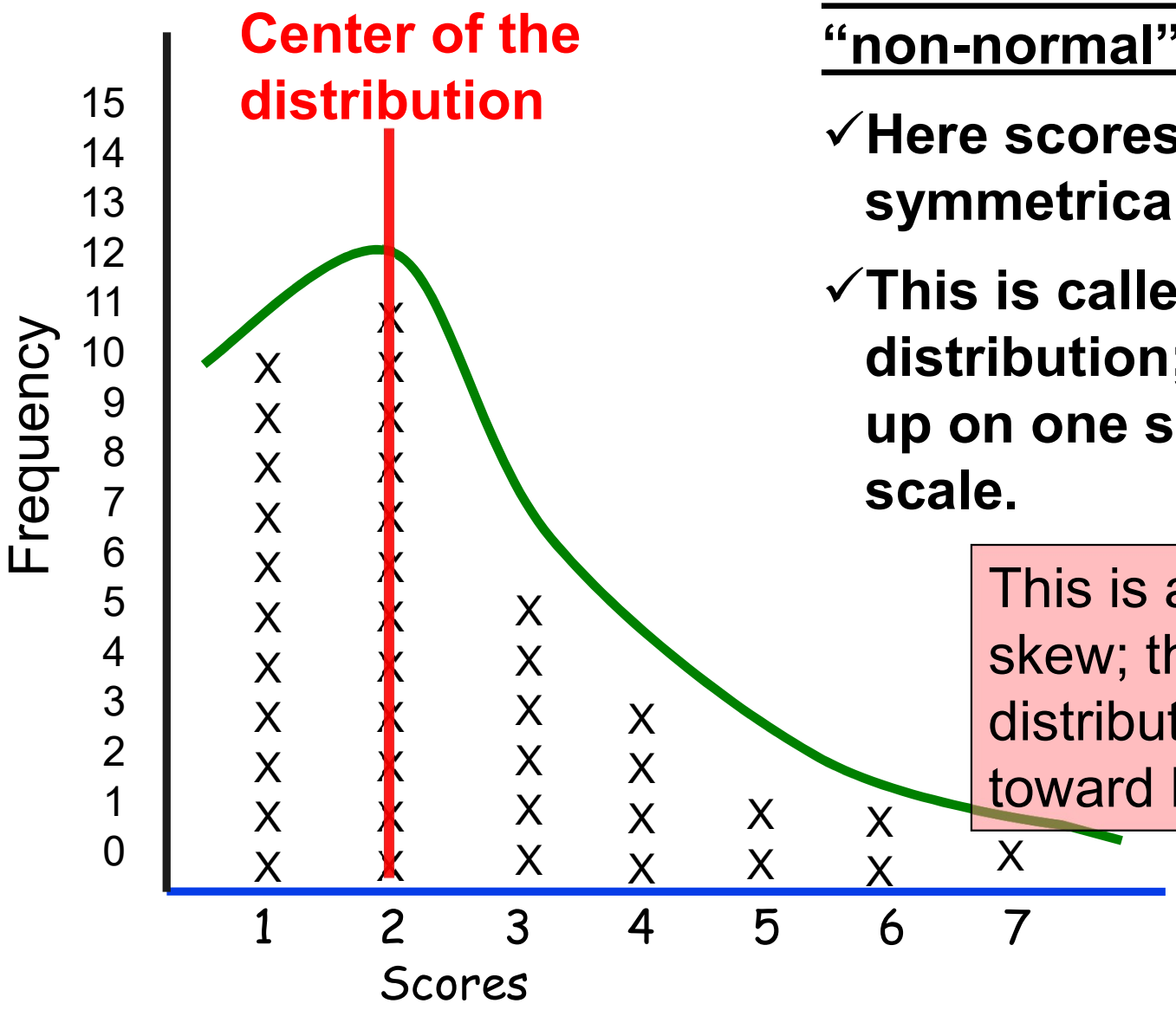


Less "normal" distributions

Score distributions may depart from the Bell Curve

- ✓Scores here are still symmetrical
- ✓The variance is "flat" or irregular





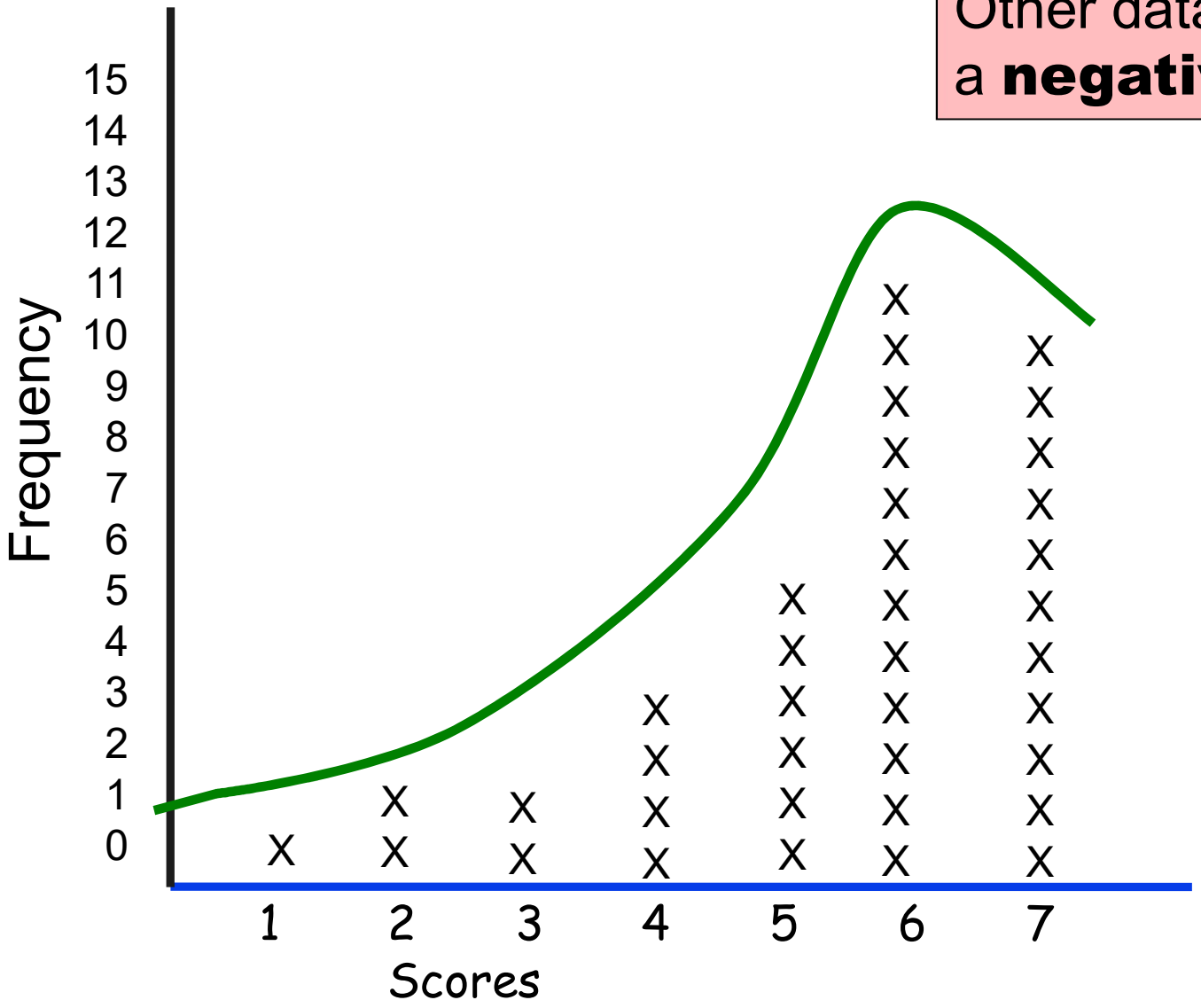
Distributions may be very "non-normal"

- ✓Here scores are not symmetrical
- ✓This is called a "skewed" distribution; scores load up on one side of the scale.

This is a positive skew; the "tail" of the distribution goes toward higher values

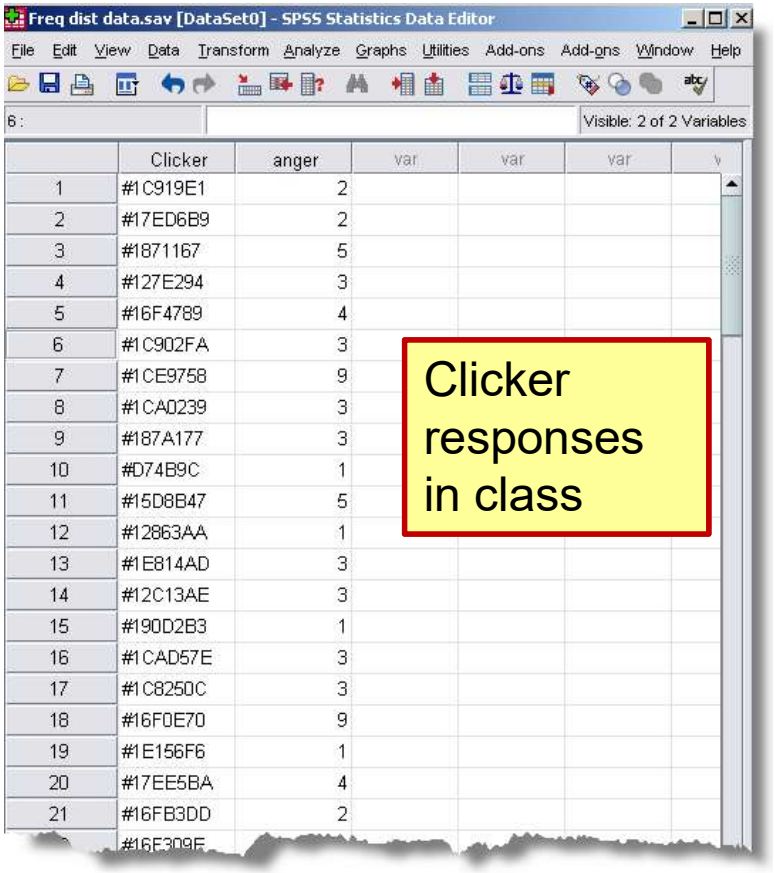
Less "normal" distributions

Other data may show a **negative** skew.



Descriptive / Univariate Statistics

1. Gather raw attitude data:


Freq dist data.sav [DataSet0] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Add-ons Window Help
6 :
Visible: 2 of 2 Variables

	Clicker	anger	var	var	var
1	#1C919E1	2			
2	#17ED6B9	2			
3	#1871167	5			
4	#127E294	3			
5	#16F4789	4			
6	#1C902FA	3			
7	#1CE9758	9			
8	#1CA0239	3			
9	#187A177	3			
10	#D74B9C	1			
11	#15D8B47	5			
12	#12863AA	1			
13	#1E814AD	3			
14	#12C13AE	3			
15	#190D2B3	1			
16	#1CAD57E	3			
17	#1C8250C	3			
18	#16F0E70	9			
19	#1E156F6	1			
20	#17EE5BA	4			
21	#16FB3DD	2			

Clicker responses in class

2. Show the number of students at each level of the variable “angry”…

anger How angry are you at investment banks?

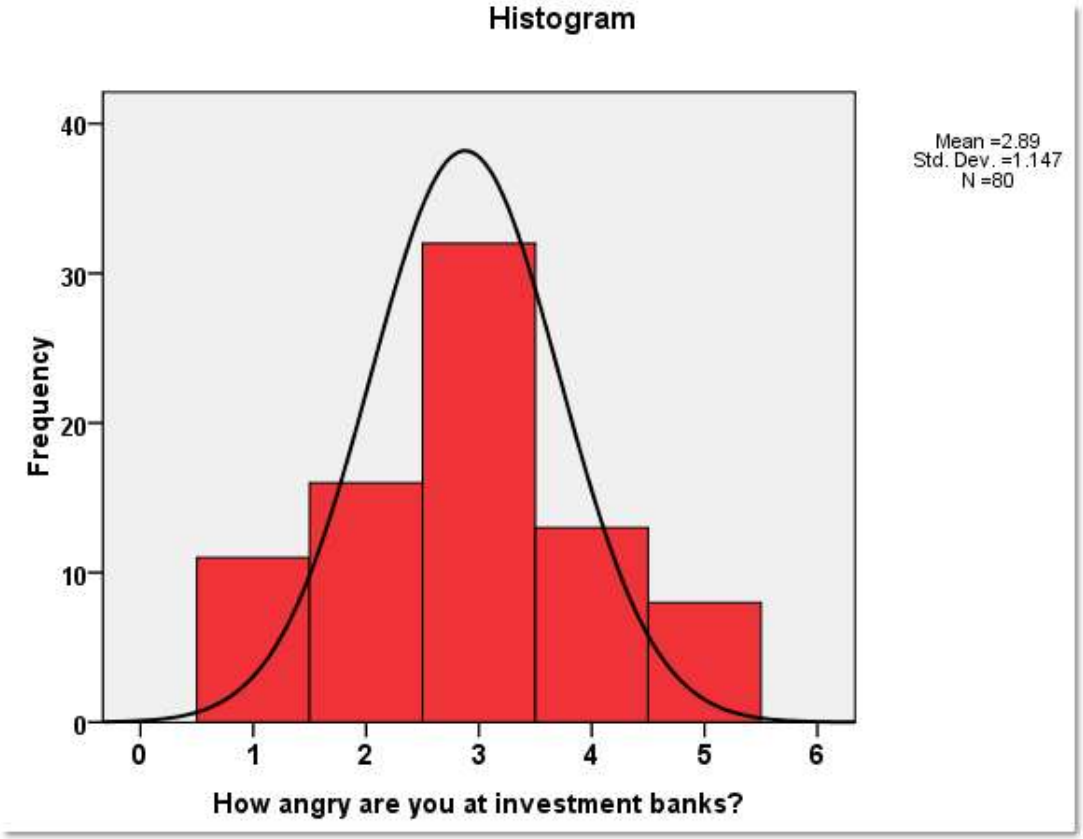
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Not at all	11	12.6	13.8	13.8
	2 A little	16	18.4	20.0	33.8
	3 Moderate	32	36.8	40.0	73.8
	4 A lot	13	14.9	16.3	90.0
	5 Extremely	8	9.2	10.0	100.0
	Total	80	92.0	100.0	
Missing	9	7	8.0		
Total		87	100.0		

Descriptive statistics example

3. Compile the descriptive statistics:

N	Valid	87
	Missing	0
Mean		3.38
Std. Error of Mean		.215
Median		3.00
Mode		3
Std. Deviation		2.001
Variance		4.006
Skewness		1.616
Std. Error of Skewness		.258
Kurtosis		2.685
Std. Error of Kurtosis		.511
Range		8
Minimum		1
Maximum		9
Sum		294

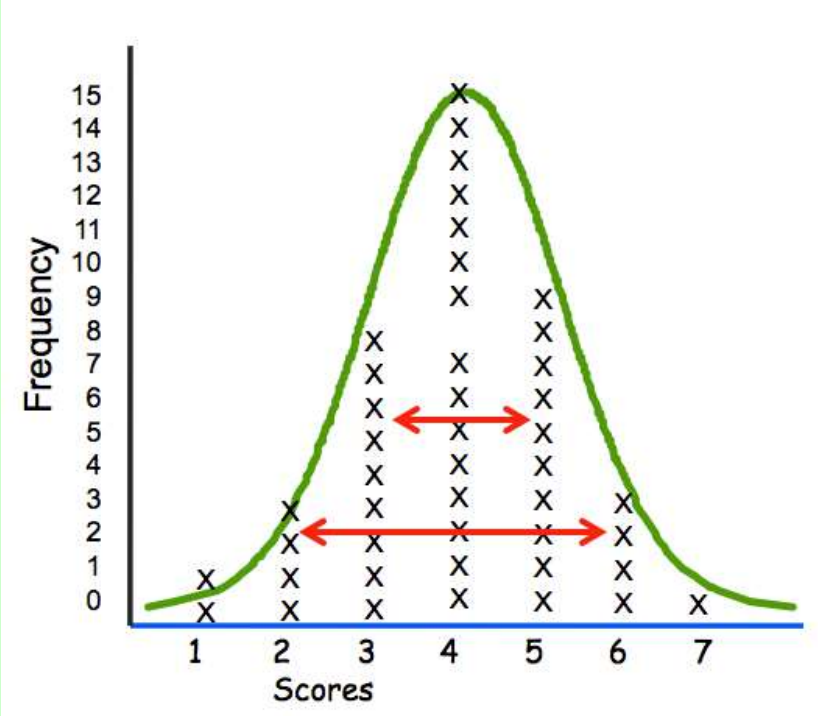
4. Show the data as a frequency distribution…



Summary: Introduction & number scales

Summary

- ✓ **Numbers are crucial to science**
  - Operational definitions
  - Communication
  - Statistical analyses
- ✓ **Types of numerical scales**
  - Ratio
  - Interval
  - Ordinal
  - Categorical
- ✓ **Numbers fall into regular frequency distributions**



Research questions, hypotheses & designs

- Using numbers in science
- Number scales & frequency distributions
- **Central Tendency: Mode, Median, Mean**

Images from the FACE RESEARCH LAB, Ben Jones and Lisa DeBruine , School of Psychology, University of Aberdeen.

- ✓ **Central tendency** – the general "drift" in a set of scores or values – can reflect any physical or ψ process or numerical scale.
- ✓ Central tendency reduces the **variance** in a sample of values to one core value.
- ✓ The following is an example from the FACE RESEARCH LAB; Ben Jones and Lisa DeBruine, School of Psychology, University of Aberdeen.

What is Central Tendency (Mean or average…)

Variance in the faces of the different lab members.

The average (”**Mean**” or **M**) face of the lab members

We characterize the general trend or character of data using two key statistics:

1. Central tendency or general "drift" of the scores.

- Mode → most common score
- Median → middle of the distribution
- Mean → average score

2. Variance: how diverse the scores are (how much *vary* from each other).

- Range → …from the highest to lowest score
- Standard deviation → "average" amount the scores vary from the Mean score

Most frequent score in the distribution.

Example: scores = 15, 20, 21, 20, 36,15, 25,15,12

- Show scores as a frequency distribution

score	frequency	% of cases
12	1	11%
15	3	33%
20	2	22%
21	1	11%
25	1	11%
36	1	11%

- 15 is most common, and is considered the mode.

Characteristics of the Mode

score	frequency	% of cases
12	1	11%
15	3	33%
20	2	22%
21	1	11%
25	1	11%
36	1	11%

- **Mode is used for all numerical scales, *particularly categorical***

  E.g., the class has 56% females & 44% males…

- **Insensitive to extreme values or range of scores**

  If we add a score of 1,000 the mode stays the same.

Characteristics of the Mode

score	frequency	% of cases
12	1	10%
15	3	30%
20	2	20%
21	1	10%
25	1	10%
36	1	10%
1,000	1	10%

- **Mode is used for all numerical scales, *particularly categorical***

  E.g., the class has 56% females & 44% males…

- **Insensitive to extreme values or range of scores**

  If we add a score of 1,000 the mode stays the same.

Characteristics of the Mode

score	frequency	% of cases
12	1	10%
15	3	30%
20	2	20%
21	1	10%
25	1	10%
36	1	10%
1,000	1	10%

- **Mode is used for all numerical scales, *particularly categorical***

  E.g., the class has 56% females & 44% males…

- **Insensitive to extreme values or range of scores**

  If we add a score of 1,000 the mode stays the same.

- **Unstable; sensitive to small shifts in numbers of cases**

Characteristics of the Mode

score	frequency	% of cases
12	1	8%
15	3	24%
20	2	16%
21	1	8%
25	1	8%
36	4	31%
1,000	1	8%

Values are rounded

- **Mode is used for all numerical scales, *particularly categorical***

  E.g., the class has 56% females & 44% males…

- **Insensitive to extreme values or range of scores**

  If we add a score of 1,000 the mode stays the same.

- **Unstable; sensitive to small shifts in numbers of cases**

  If we add three more scores of 36 the mode changes a lot.

Median

Mid-point of a distribution of scores:

- Half are above, half are below
- Only for continuous (interval or ratio) scales

- List all the scores in numerical order

score	frequency	% of cases
12	1	11%
15	3	33%
20	2	22%
21	1	11%
25	1	11%
36	1	11%

12
15
15
15
20
20
21
25
36

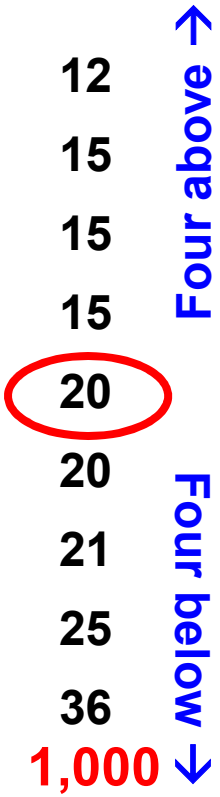
Median

Mid-point of a distribution of scores:

- Half are above, half are below
- Only for continuous (interval or ratio) scales

- List all the scores in numerical order
- Locate the score in the center of the order

  For an even number of scores the Median is the average of the middle two scores
- The Mode = 15, whereas Median = 20
  - For this distribution the median best reflects the center of the distribution
- Median is not sensitive to extreme values
  - Adding '1,000' would not matter; '20' would still be the center of the distribution.



The "average" score in sample; $\sum x / n$

Most common measure of central tendency

- **Total all scores: 12+15+20+21+20+36+15+25+15 = 179**
- **Divide by "n" of scores: 179 / 9 = 19.9 (*round to* 20)**

- **Characteristics:**
  - **Good for Ratio or interval scales**
  - **Sensitive to all observed values**
  - **Highly stable; with larger n is insensitive to subtle changes in values**
  - **Can be highly sensitive to extreme values.**

    12+15+20+21+20+36+15+25+15+1000 = 1,179 / 10 = 118

Some basic statistical notation:

X **Score** on one variable for one participant

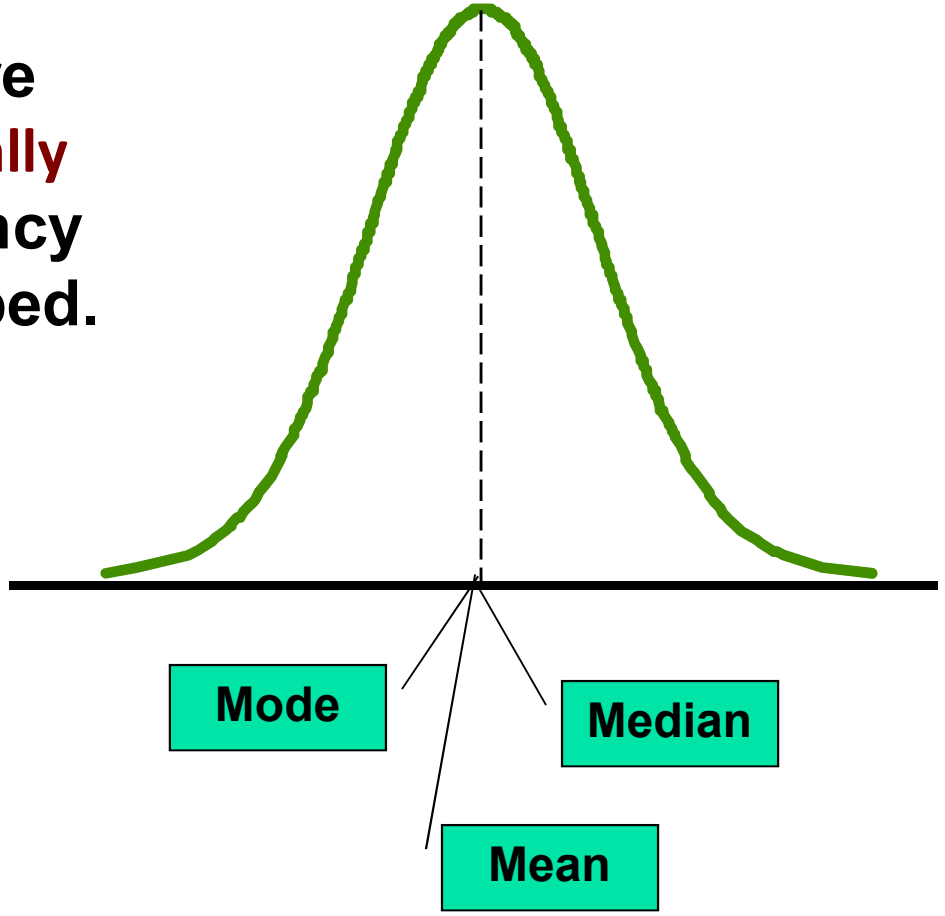
n **Number** of scores in the sample

Σ **Sum** of a set of scores

M **Mean**; sum of scores divided by n of scores: $\sum x / n$

Central tendency: Normal Distributions

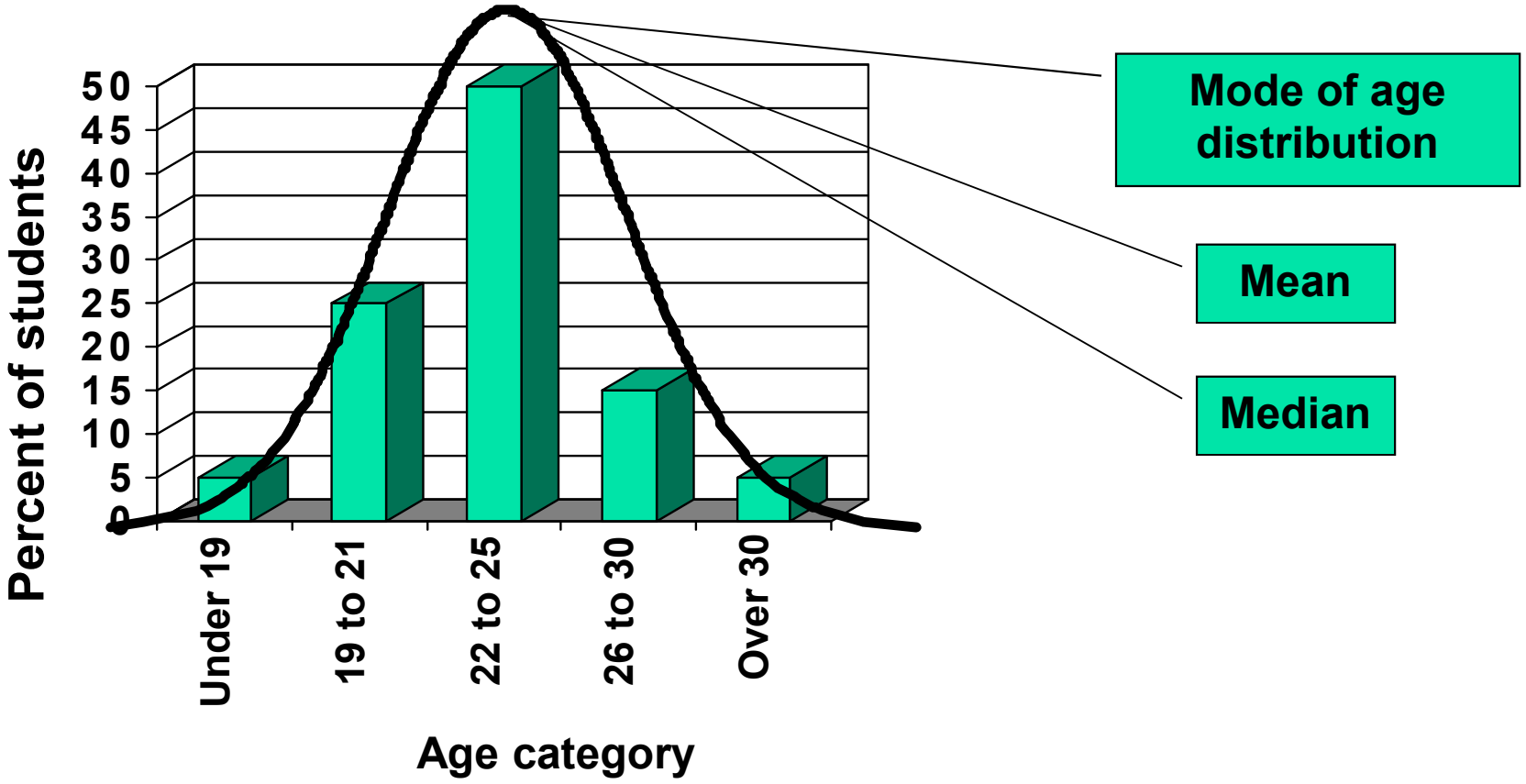
- Many variables in nature (and science) are normally distributed; their frequency distribution is bell shaped.



- For a normal distribution the *mean, mode*, and *median* are all same -- the center of the distribution

The distribution of student ages

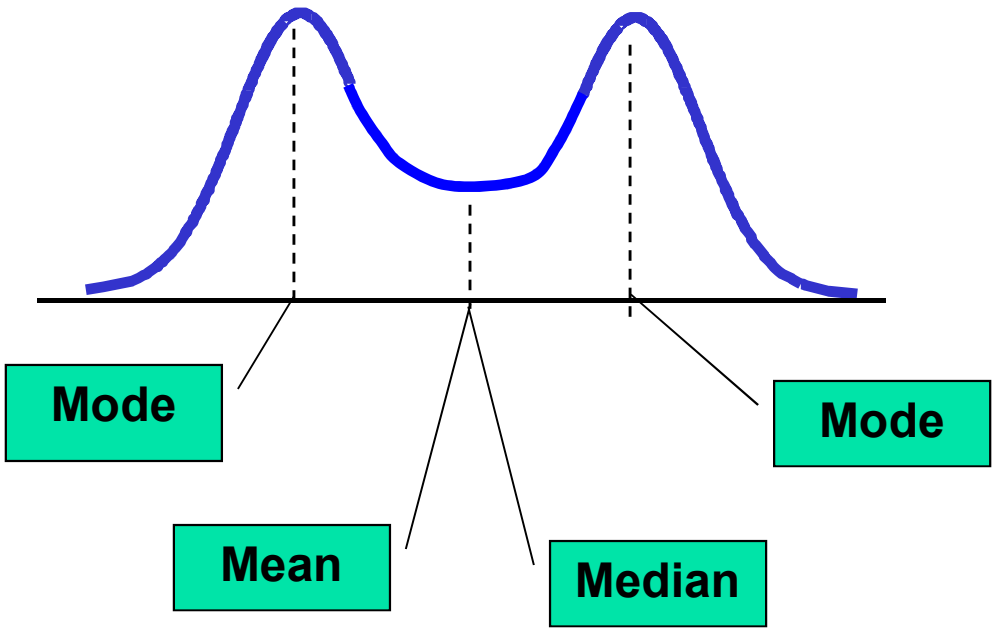
Age is a good example of a variable that is *normally distributed*



Central tendency: Bimodal Distributions

A bimodal distribution has two *modes*, at the outsides of the distribution.

The *mean* & *median* are similar, at the center.



Common examples are:

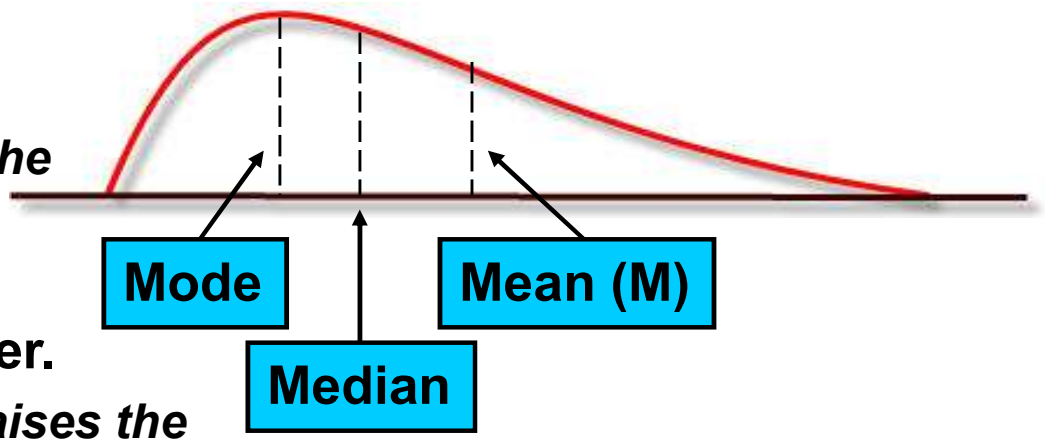
- ✓ **Highly polarized political attitudes *(i.e., where few are "in the middle").***
- ✓ **Some personality variables..**

Those who fear and loathe statistics v. those who love statistics and are popular, mentally healthy, and happy,

A skewed distribution has extreme scores in one direction.

The extreme scores make the median higher than the mode.

(The high scores to the right move the 50% point that direction…).



The Mean gets pulled even higher.

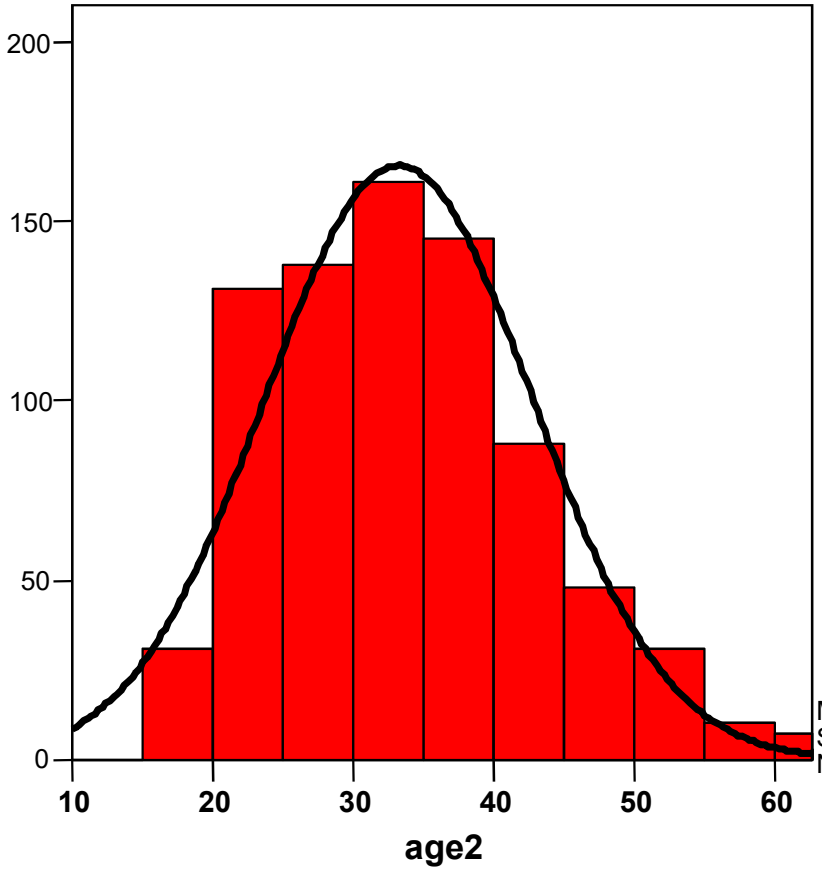
(Adding in some very high scores raises the average…).

Common examples:

- Behaviors such as alcohol or drug use:
  - Most people use none or moderate
  - A diminishing number use higher levels
- Demographic variables such as income

Age, Chicago community sample

N	Valid	793
	Missing	24
Mean		33.1967
Median		32.0000
Mode		30.00
Std. Deviation		9.54557
Variance		91.118
Skewness		.633
Range		50.00



Measures of Central Tendency: A normal distribution

- Scores for age from a large community sample form a largely symmetrical distribution.
- The Mean, Median, and mode are similar.
- Any measure of central tendency well represents the data.

Local examples of distributions, 1

How many packs of cigarettes do you smoke each week?

A = I do not smoke

B= 1 pack or less

C = 2 packs or so

D = 3 to 5 packs

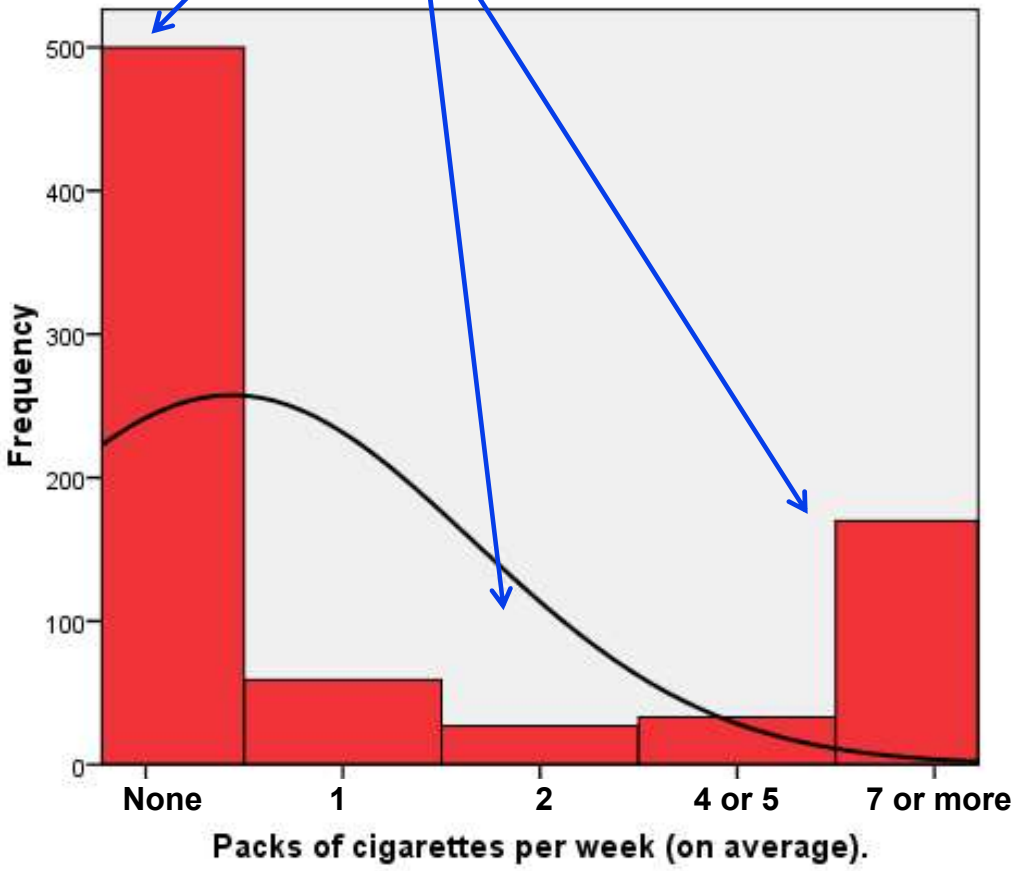
E = A pack a day or more

Shutterstock.com

Cigarette smoking forms a bimodal distribution

- People do not smoke at all…
- …or smoke a lot.
- Few are "light" smokers.

- Tobacco use is a good example of a bimodal distribution.



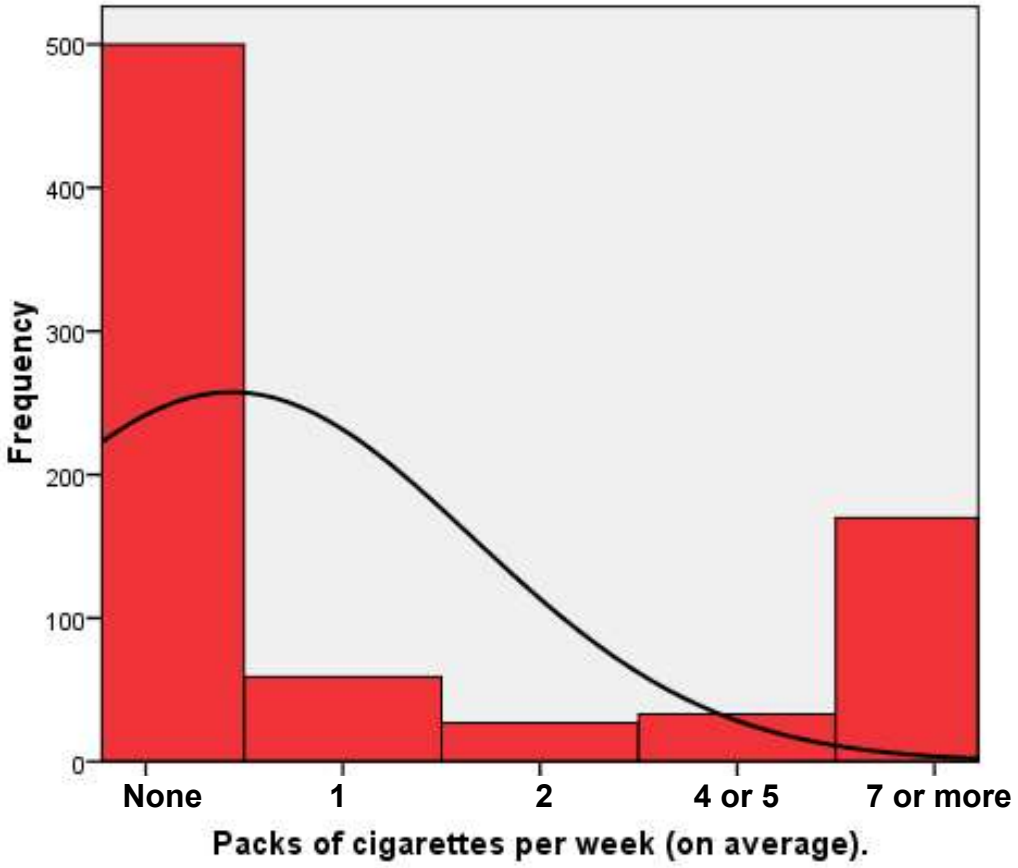
Statistics

Packs of cigarettes per week

N	Valid	789
	Missing	28
Mean		1.1305
Std. Error of Mean		.05915
Median		.0000
Mode		.00
Std. Deviation		1.66
Variance		2.761
Skewness		.978
Std. Error of Skewness		.087
Kurtosis		-.871
Std. Error of Kurtosis		.174
Range		4.00
Minimum		.00
Maximum		4.00
Sum		892.00

Cigarette smoking forms a bimodal distribution

- People do not smoke at all…
- …or smoke a lot.
- Few are "light" smokers.
- Mean = 1.1 packs per week clearly misrepresents the data.
- The Mode or Median are better indicators.



Statistics

Packs of cigarettes per week

N	Valid	789
	Missing	28
Mean		1.1305
Std. Error of Mean		.05915
Median		.0000
Mode		.00
Std. Deviation		1.66
Variance		2.761
Skewness		.978
Std. Error of Skewness		.087
Kurtosis		-.871
Std. Error of Kurtosis		.174
Range		4.00
Minimum		.00
Maximum		4.00
Sum		892.00

Local examples of distributions, 2b

During the <u>last week</u> on how many days did you have at least one drink of alcohol?

A = 0

B= 1

C = 2 or 3

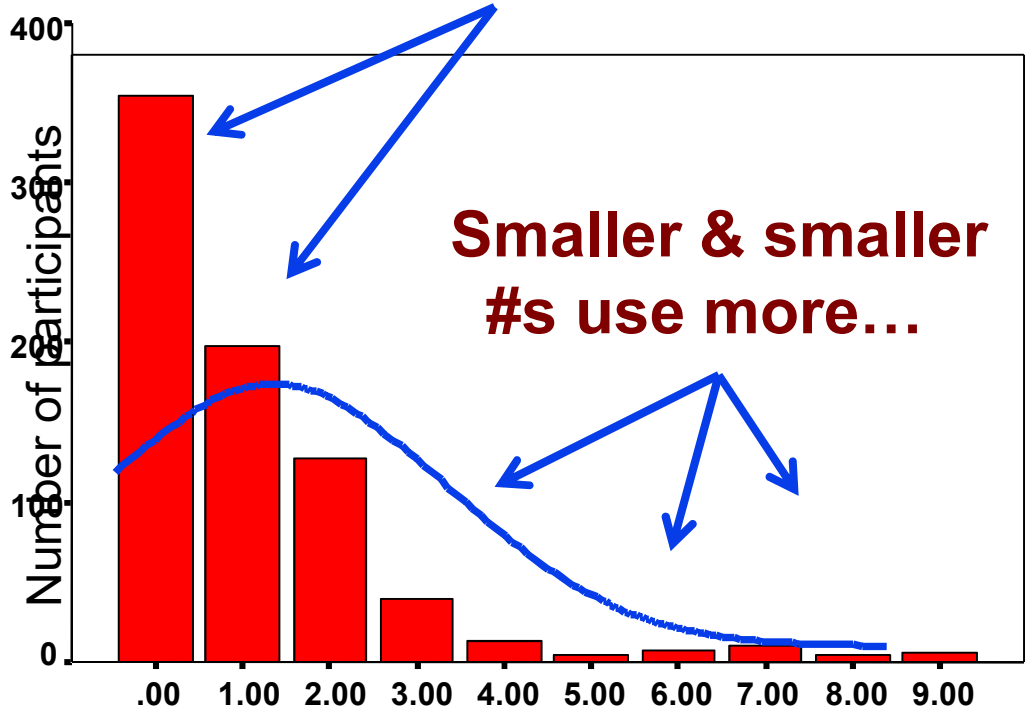
D = 4 or 5

E = 6 or 7

Shutterstock.com

Drug or alcohol use exemplify positively skewed data

Most people use no or few substances



This strong positive skew is reflected in a skewness statistic

Statistics

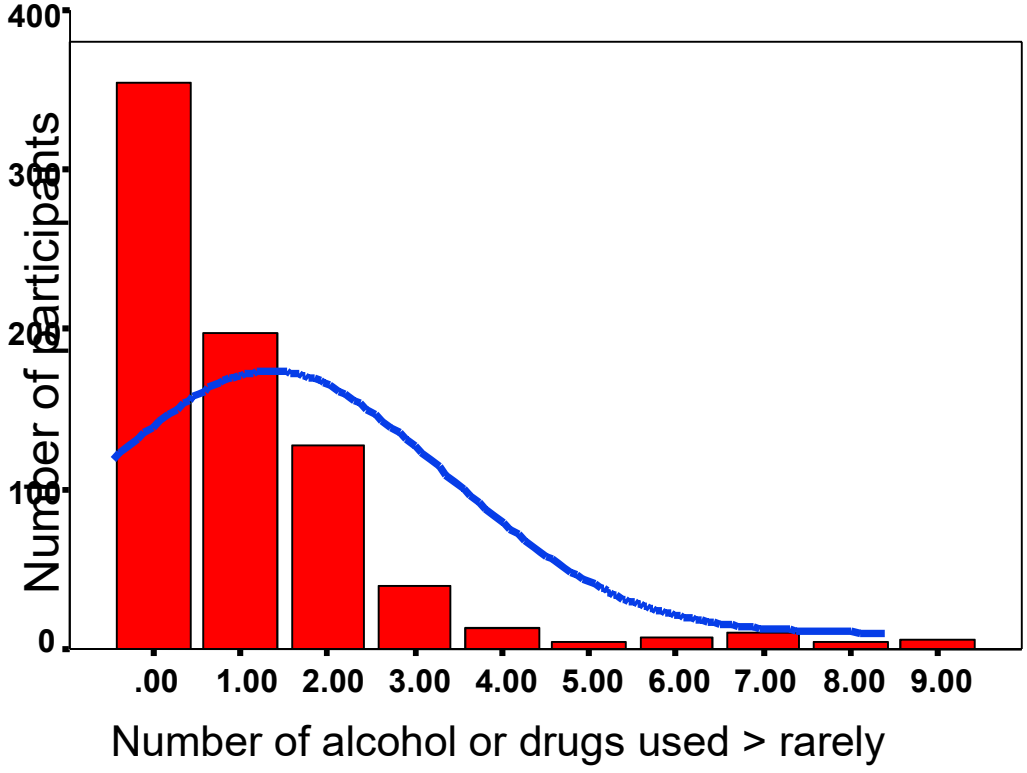
Number of alcohol or drugs used > rarely

N	Valid	766
	Missing	51
Mean		1.11
Median		1.00
Mode		.00
Std. Deviation		1.59
Variance		2.54
Skewness		2.46
Minimum		.00
Maximum		9.00

(Data from a community survey sample)

Drug or alcohol use exemplify positively skewed data

With a strong positive skew it is not clear which measure of central tendency is best.



The Mode, Median & Mean differ

Statistics

Number of alcohol or drugs used > rarely

N	Valid	766
	Missing	51
Mean		1.11
Median		1.00
Mode		.00
Std. Deviation		1.59
Variance		2.54
Skewness		2.46
Minimum		.00
Maximum		9.00

The Mode or median may be used for central tendency here.

(Data from a community survey sample)

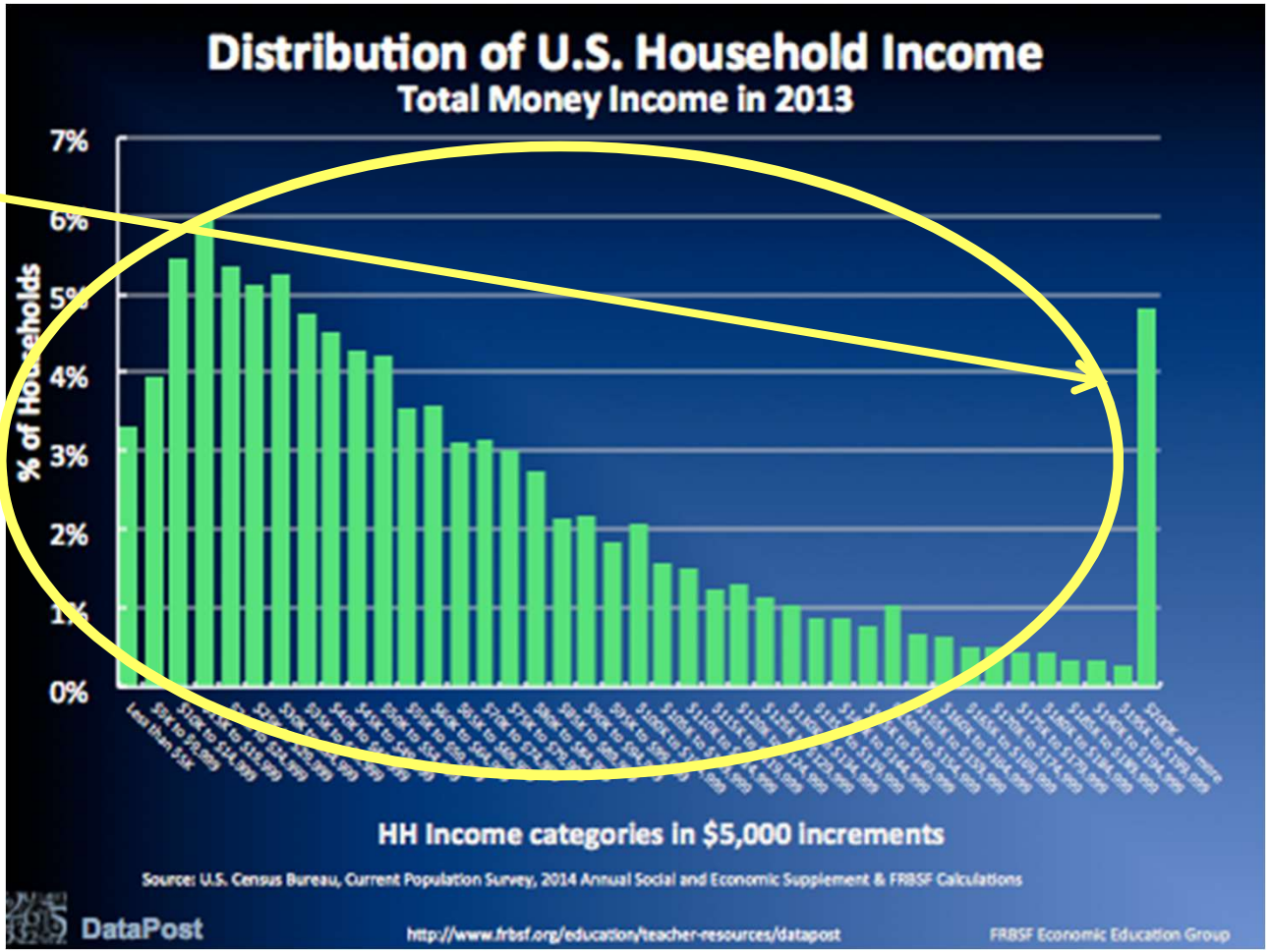
Skewed distributions: income.

Income represents one of the most skewed distributions imaginable.

The extreme incomes of the top 0.5% or so substantially distort the distribution .

Incomes from 200K to hundreds of millions per year

99% of all incomes.



Skewed distributions: income.

Income represents one of the most skewed distributions imaginable.

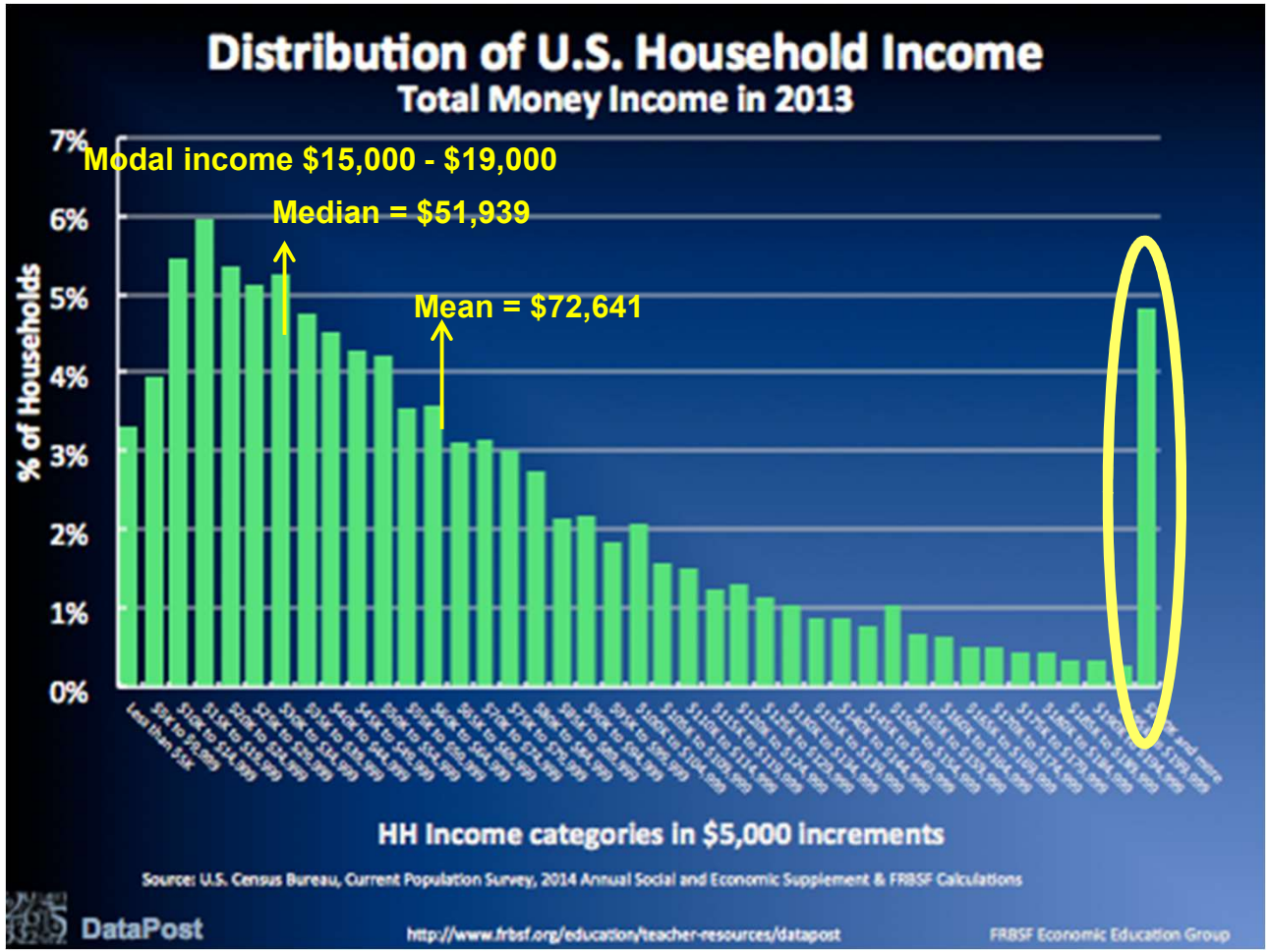
The extreme incomes of the top 0.5% or so substantially distort the distribution .

Modal family income is under $20K.

Mean income is over 4 times that of the mode.

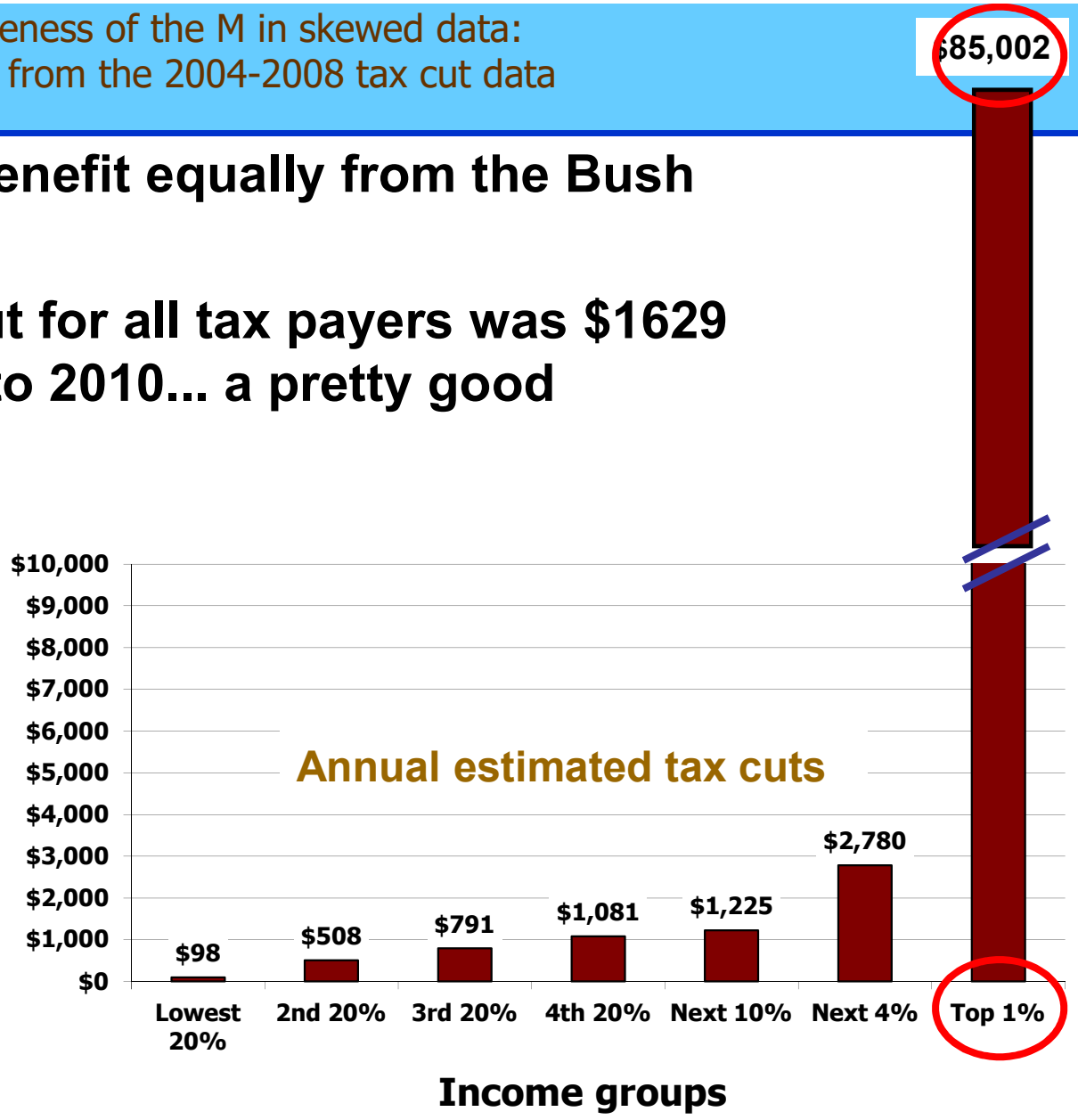
The Mean is "pulled up" by the very high values at the top of the scale.

In a distribution such as this we typically accept the Median as the more representative value.



Deceptiveness of the M in skewed data: Example from the 2004-2008 tax cut data

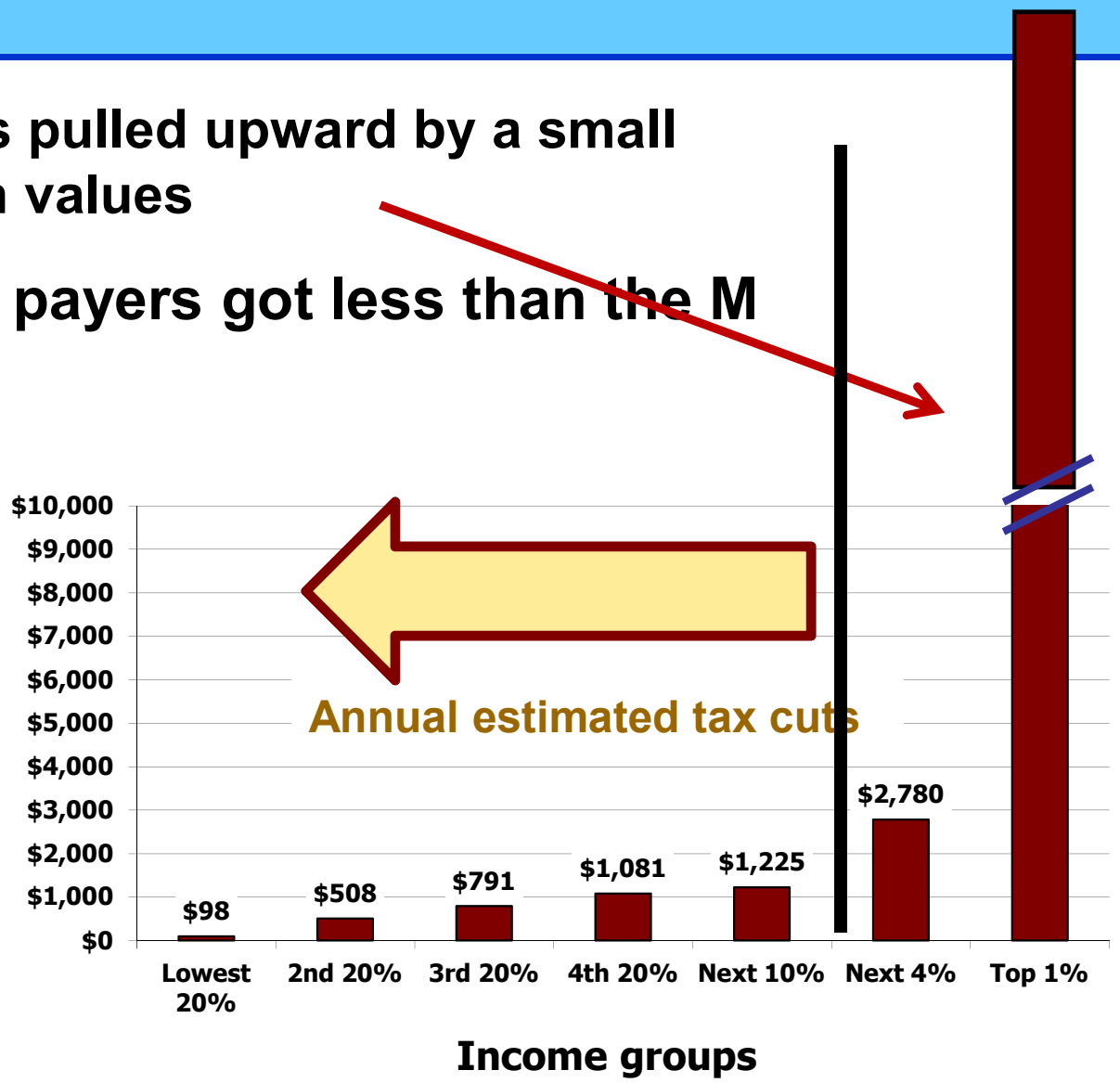
- **Did all Americans benefit equally from the Bush era tax cuts?**
- **The M annual tax cut for all tax payers was $1629 per year from 2001 to 2010... a pretty good number.**
  - **The extreme positive skew in incomes & tax legislation make this highly deceptive.**



Source: Citizens for Tax Justice, 6/12/06; click here

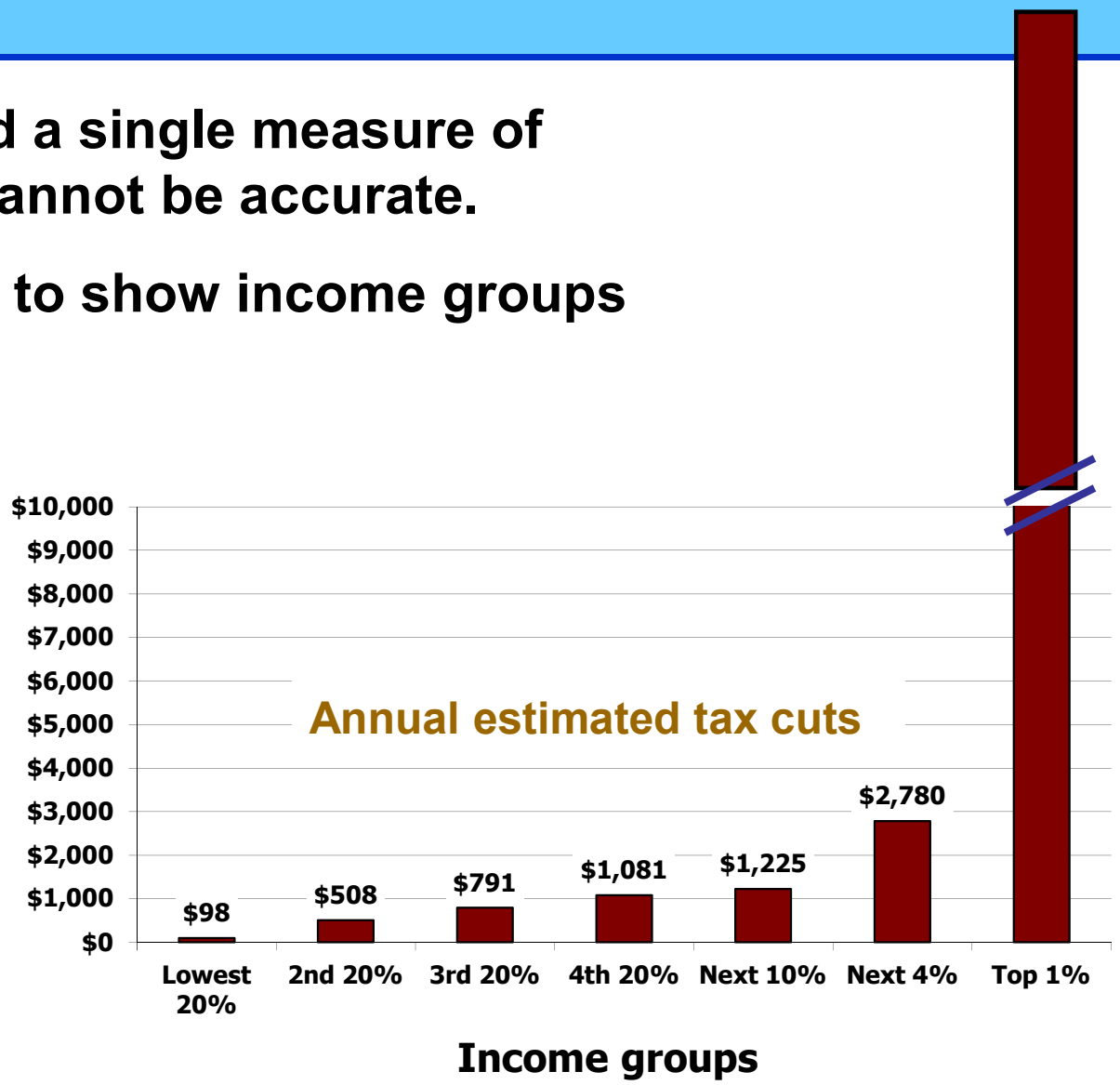
Deceptive Means in skewed data

- The M for tax cuts is pulled upward by a small number of very high values
- Almost 90% of tax payers got less than the M tax cut of $1,629

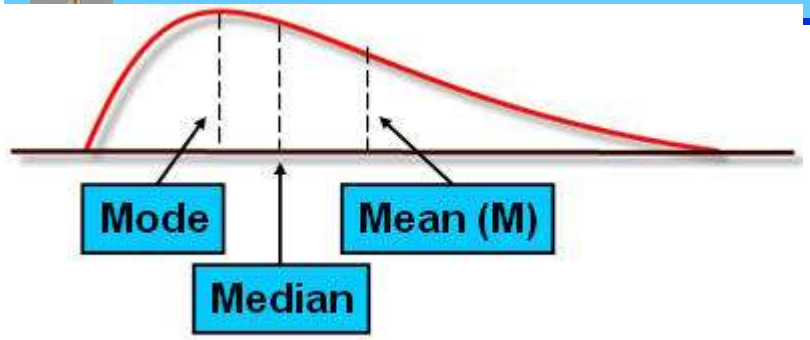


Deceptive Means in skewed data

- **In data this skewed a single measure of central tendency cannot be accurate.**
- **It is more accurate to show income groups separately.**
  - **The bottom 66% of tax payers got M = $461 in cuts,**
  - **Top 10% M = $10,203**

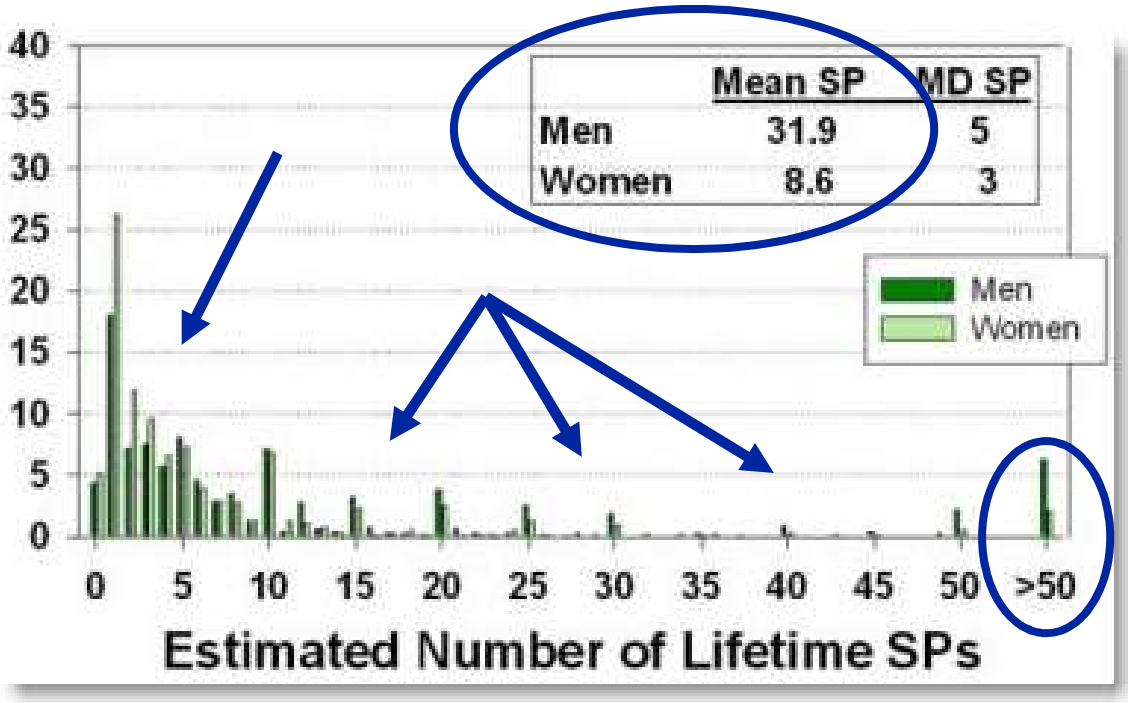


Mean v. Median in skewed data



The difference between the Mode, Median & Mean in skewed data can be crucial: Example from men's v. women's number of sex partners.

- M number of partners is much higher for men than for women.
- However, both men & women have highly skewed distributions:
  - Most at the low end
  - A long tail to the right
  - A small mode at very high numbers



From Phys.org. Click for original article.

Expressing the data as Ms greatly exaggerates differences between men & women.

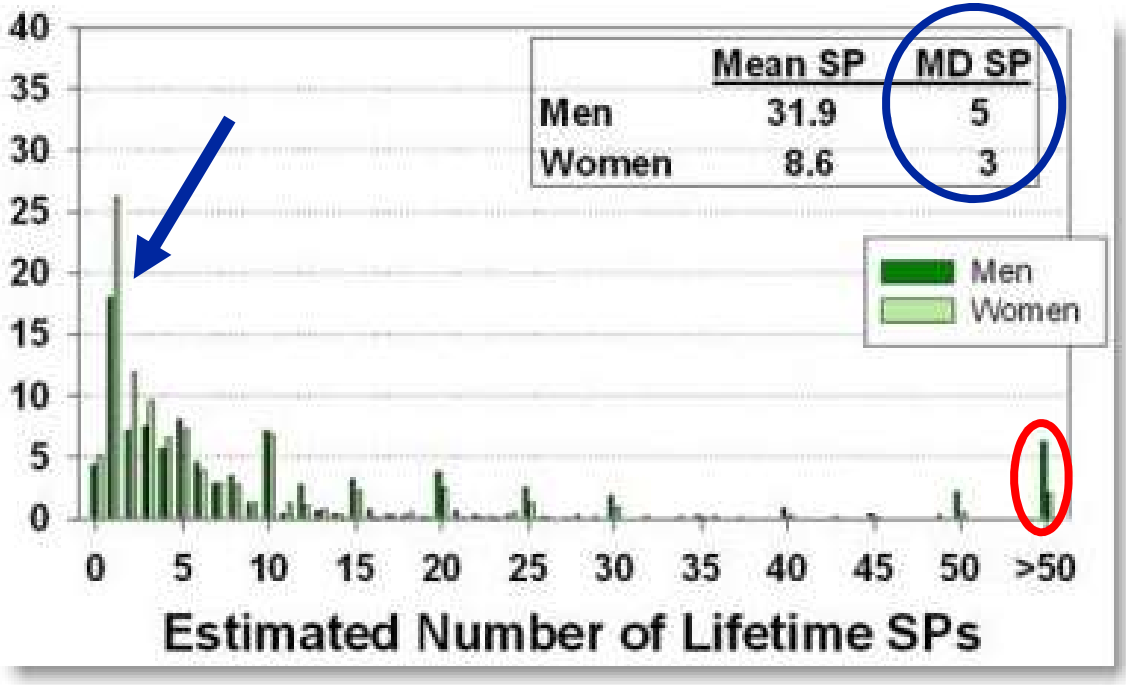
For both genders:
- ✓mode = 1,
- ✓medians are similar (5 v. 3),

This suggests only modest gender differences.

Men's data are *far* more skewed – some men report A LOT of partners – their M is very high (6 times their median…).

The genders are similar, except for a few men at the very top of the scale, who pull their M upwards.

Simply taking the Ms at face value would be deceiving.



Bottom line:

- **Many natural processes or variables show a** (more or less) **normal distribution…**
  - **Age, Height, IQ, most attitude scales…**
- **Many important behavioral variables have a highly skewed distribution.**
- **When a distribution is normal the *Mode* = *Mean* = *Median*, all at the center of the distribution.**
- **When a distribution is bimodal or skewed the M can be deceptive.**
  - **Mode** → **Best for categorical data or a simple summary**
  - **Median** → **Best for highly skewed data**
  - **Mean** → **best for normally distributed data.**

Summary: Introduction & number scales

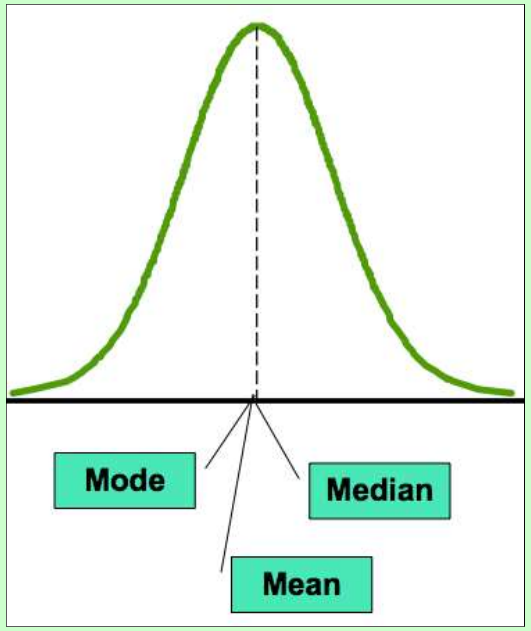
Summary

✓ Primary measures of Central Tendency

- Mode → Most common value
- Median → Middle of the distribution
- Mean → Average

✓ Types of frequency distributions

- "Normal" → Best characterized by the Mean



Summary: Introduction & number scales

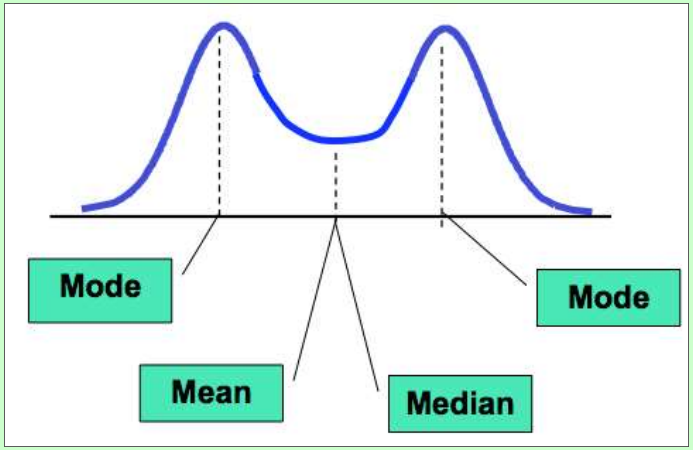
Summary

✓ Primary measures of Central Tendency

- Mode → Most common value
- Median → Middle of the distribution
- Mean → Average

✓ Types of frequency distributions

- "Normal"
- Bimodal → Best characterized by the Modes



Summary: Introduction & number scales

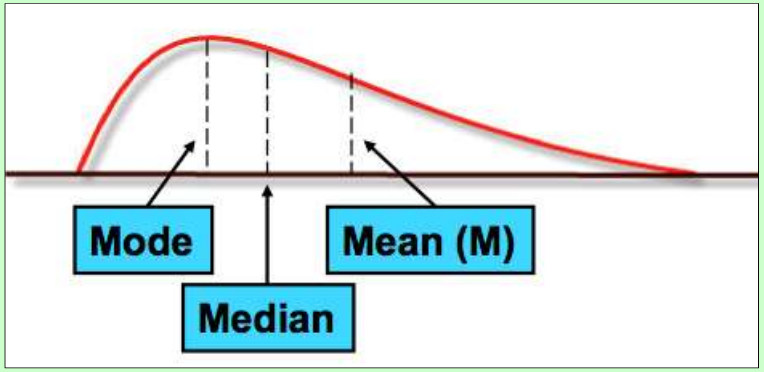
Summary

✓ Primary measures of Central Tendency

- Mode → Most common value
- Median → Middle of the distribution
- Mean → Average

✓ Types of frequency distributions

- "Normal"
- Bimodal
- Skewed → Best characterized by the Median or mode



In the next module we will add measures of Variance to Central Tendency, and combine them into the Z score.